Argument & Computation 15 (2024) 49–107
DOI 10.3233/AAC-220011
IOS Press

Information-seeking dialogue for explainable artificial intelligence: Modelling and analytics

Ilia Stepin [a,c,*], Katarzyna Budzynska [b], Alejandro Catala [a,c], Martín Pereira-Fariña [d] and Jose M. Alonso-Moral [a,c]

[a] *Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Rúa de Jenaro de la Fuente Domínguez s/n, 15782 Santiago de Compostela, A Coruña, Spain*
E-mails: ilia.stepin@usc.es, alejandro.catala@usc.es, josemaria.alonso.moral@usc.es
[b] *Laboratory of The New Ethos, Warsaw University of Technology, plac Politechniki 1, 00-661, Warsaw, Poland*
E-mail: katarzyna.budzynska@gmail.com
[c] *Departamento de Electrónica e Computación, Universidade de Santiago de Compostela, Rúa Lope Gómez de Marzoa, s/n, 15782 Santiago de Compostela, A Coruña, Spain*
[d] *Departamento de Filosofía e Antropoloxía, Universidade de Santiago de Compostela, Plaza de Mazarelos s/n, 15705 Santiago de Compostela, A Coruña, Spain*
E-mail: martin.pereira@usc.es

Abstract. Explainable artificial intelligence has become a vitally important research field aiming, among other tasks, to justify predictions made by intelligent classifiers automatically learned from data. Importantly, efficiency of automated explanations may be undermined if the end user does not have sufficient domain knowledge or lacks information about the data used for training. To address the issue of effective explanation communication, we propose a novel information-seeking explanatory dialogue game following the most recent requirements to automatically generated explanations. Further, we generalise our dialogue model in form of an explanatory dialogue grammar which makes it applicable to interpretable rule-based classifiers that are enhanced with the capability to provide textual explanations. Finally, we carry out an exploratory user study to validate the corresponding dialogue protocol and analyse the experimental results using insights from process mining and argument analytics. A high number of requests for alternative explanations testifies the need for ensuring diversity in the context of automated explanations.

Keywords: Explainable Artificial Intelligence, information-seeking dialogue game, explanation locutions, counterfactual explanation, process mining analytics, argument analytics

1. Introduction

Explainability in the context of Artificial Intelligence (AI) has long attracted attention of researchers from computer science [57] and argumentation [21]. The first explanation generation methods turned up

*Corresponding author. Tel.: +34 8818 16394; E-mail: ilia.stepin@usc.es.

in the 1980s along with the so-called Expert Systems [74]. More precisely, the first explainers addressed the challenge of explaining the output of expert systems and logic programs [7], which eventually led to the emergence of the research field that we now call Computational Argumentation. Recent years have witnessed a new boost of interest in developing eXplainable AI (XAI), as novel machine learning (ML) algorithms produce highly accurate yet oftentimes poorly explainable predictions [1]. As defined at present, XAI aims to (1) generate explainable models preserving a high level of accuracy and (2) enable the end user, e.g., a client of a bank or a patient of a hospital, with the opportunity to understand, trust, and manage the given AI-based systems [2,29] (e.g., querying a bank loan management system to identify reasons for the loan application being rejected or a hospital information system to receive treatment-related recommendations).

The obscure nature of the underlying reasoning of the state-of-the-art predictive algorithms has given way to the so-called "right to explanation" [80]. The corresponding legal regulations are being increasingly adopted worldwide [87]. For example, the European Union (EU)'s General Data Protection Regulation (GDPR) acknowledges the right of the user "not to be subject to a decision evaluating personal aspects relating to him or her which is based solely on automated processing and which produces adverse legal effects concerning, or significantly affects, him or her" [51]. In addition, current EU's legal regulations in, for example, the financial domain require that algorithmic transparency be provided for automatic trading techniques (see the Directive 2014/65/EU on Markets in Financial Instruments, commonly known as MiFID II [52] for details). Being a controversial topic of primary importance for numerous stakeholders, its juridical basis is constantly updated. Thus, the newly proposed EU's AI Act (AIA) [53] establishes a taxonomy of AI-based systems and requires that high-risk AI applications offer explanations for their decisions or recommendations to their end users.

In order to mitigate algorithmic transparency issues of the state-of-the-art AI algorithms, a use of interpretable models is advised [59]. Interpretable rule-based models (such as, e.g., decision trees (DT) or decision rules) are known to provide user-friendly explanations [47]. Remarkably, DTs can be used as part of more complex model-agnostic explainers that are able to justify predictions of other arbitrary classifiers if they are, for example, trained on a local synthetically generated neighbourhood around the test instance [28]. Despite the fact that only few XAI frameworks offer explanations in natural language [12], DTs have also been shown to be a powerful tool for communicating textual explanations to end users, e.g., by engaging the user in an explanatory dialogue [70,79].

Explanations are claimed to have to necessarily be embedded in a dialogical interaction so that the end user is able to challenge the aspects of an explanation that have not been understood [63]. For illustrative purposes, let us consider a beer style classification problem (see Table 1 for details). Given a number of predefined classes (i.e., beer styles) and an instance of beer, the classifier (System) makes a prediction on what beer style the test instance is (move m_1). The end user (User) engages in a dialogue with the System to inquire an explanation for the given prediction (m_2). Then, the System provides the User with a factual explanation based on the most distinctive features, pertinent to the test instance, that led to the given prediction (m_3). Then, given the factual explanation, the User may want to clarify specific features (e.g., bitterness) that the offered explanation contains (m_4). In response, the System offers a definition for the requested feature (m_5). Alternatively, the User may inquire the System to explain why the given test instance is not predicted to be of another class (m_6, m_{10}). The pieces of such counterfactual explanations offered in response suggest minimal changes to the input so that its classification changes in the desired manner (m_7, m_{11}). Given the initial explanation, the User may want to specify a more precise description of the given feature (m_8). Then, the System may present detailed information over the corresponding numerical interval (subject to availability) over which the feature is defined (m_9). Further,

Table 1
An illustrative explanatory dialogue

Move	*Agent*	*Utterance*
m_1	System	The present beer is Blanche.
m_2	User	Could you explain why you think so?
m_3	System	It is Blanche because colour is pale and bitterness is low.
m_4	User	What do you mean by bitterness?
m_5	System	Bitterness is a harsh, acrid taste that is one of the four basic taste sensations.
m_6	User	But is the beer not stout?
m_7	System	It would be Stout if colour were black and bitterness were high.
m_8	User	Could you specify how high bitterness is defined?
m_9	System	Sure! I define bitterness as high because it lies in the range from 47 to 250.
m_{10}	User	But why is it not Porter?
m_{11}	System	It would be Porter if colour were brown.
m_{12}	User	I am not quite satisfied with your explanation. Could you offer me another one?
m_{13}	System	Sure! It would be of class Porter if colour were brown and strength were session.
m_{14}	User	Ok, now I trust your prediction.
m_{15}	System	Thank you for your trust in me. Bye!

the User may disagree with the explanation offered and argue over it (m_{12}). The System should then offer an alternative explanation that would satisfy the User's needs (m_{13}). When the User is sufficiently informed about the reasons that led to the given prediction, he or she makes an informed decision on whether the System's prediction should be trusted or not (m_{14}). The explanatory dialogue ends with the System's farewell locution (m_{15}).

As follows from Table 1, we consider two types of explanations: factual and counterfactual. Assuming knowledge of the feature space, factual explanations (illustrated with move m_3 in Table 1) aim to explain the given classifier's prediction in terms of the most relevant feature values that led to that prediction. On the contrary, counterfactuals (illustrated with moves m_7, m_{11}, and m_{13} in Table 1) are post-hoc example-based explanations that suggest a minimal change in feature values to those of the given data instance so that the system's prediction changes as desired [71].

This paper introduces an explanatory dialogue game for communicating factual and counterfactual explanations for interpretable rule-based classifiers. We assume that the classifier is associated with an explainer that is capable of providing textual (rule-based) explanations. Based on the dialogue typology proposed by Walton and Krabbe [82], we model the information-seeking type of explanatory dialogue equipping it with a specific collection of locutions tailored for the aforementioned types of explanation that the user may ask the system. As a starting point, we consider the typology of dialogue moves proposed by Budzynska et al. [9]. In our work, we extend this typology of dialogue moves with a repertoire of locutions allowing for communication of factual and counterfactual explanations to enable the end user to interactively explore the explanation space. Then, we propose a context-free dialogue grammar to generalise the formal structure of the resulting dialogue model. Despite an empirically shown strong need in both factual and counterfactual explanations [41] and at least a hundred of counterfactual explanation generation methods proposed by now in the context of XAI, less than a third of these methods are evaluated in user studies [37]. To address this issue, we subsequently perform a pilot user study to evaluate the proposed dialogue model. Moreover, we analyse the collected dialogue transcripts treating instances of explanatory dialogue as processes using the state-of-the-art techniques from process mining and argument analytics [43].

As a result, we bridge the gap between ML practitioners and the argumentation community by making the following contributions:

- we model information-seeking explanatory dialogue based on the fundamental notions from the argumentation theory and apply the dialogue model in the context of XAI;
- we propose a set of original dialogue locution types that are found specifically suitable for effective communication of factual and counterfactual explanations;
- we demonstrate the explanatory utility of the proposed dialogue protocol via a human evaluation study based on three use cases for an interpretable rule-based classifier leaving open-source implementations of the dialogue game and the human evaluation toolkit available for public use;
- we suggest formal means for extending the proposed protocol to make it applicable to modelling dialogic human-machine interaction for classification tasks in other applications.

The rest of the manuscript is structured as follows. Section 2 introduces the classification problem formally and outlines the common properties of explanations claimed to be essential for explaining solutions to such a problem. In addition, we subsequently discuss possible discrepancy between automatically generated explanations and user-preferred explanations. Section 3 defines an explanatory dialogue game as an interface between an explanation generation module and the end user. Section 4 introduces essential process mining concepts and shows how we apply them to explanatory dialogue analysis. Section 5 presents the experimental settings of the human evaluation study carried out to assess the utility of the proposed dialogue protocol. Section 6 reports the experimental results obtained from the human evaluation study. Section 7 discusses the dialogue model validation results. Section 8 presents an overview of related work regarding formal explanatory dialogue models as well as recent argumentation-based techniques for explanatory dialogue modelling. Finally, we outline prospective directions for future work and conclude in Section 9.

2. Preliminaries

In this section, we first outline a definition of the classification problem and assumptions about the nature of classifiers and explainers that we are driven by (see Section 2.1 for details). Then, we formally define essential explanation-related concepts that we utilise throughout the manuscript in Section 2.2. Finally, we draw reader's attention to possible discrepancies between the user-preferred explanations and those offered to him or her by the explainer in Section 2.3.

2.1. The classification problem

As outlined in Section 1, we focus on communicating to the end user automated explanations for the output of an interpretable rule-based ML classifier. Figure 1 depicts a general architecture of the modelled explanation communication process. The System is assumed to include, at least, the following core components: an interpretable rule-based classifier, an explainer, a knowledge base, and a dataset that the classifier is trained on. The User starts the communication process by sending a classification request for a specific test instance to the System in form of the test instance's characteristics (i.e., features). The classifier is pretrained on a given dataset $X = \{x_i\}|_{i=1}^{n}$ containing n labelled instances to learn a mapping function $c : X \longrightarrow Y$ where $Y = \{y_j\}|_{j=1}^{m}$ is a discrete output variable (class), m being the number of classes.

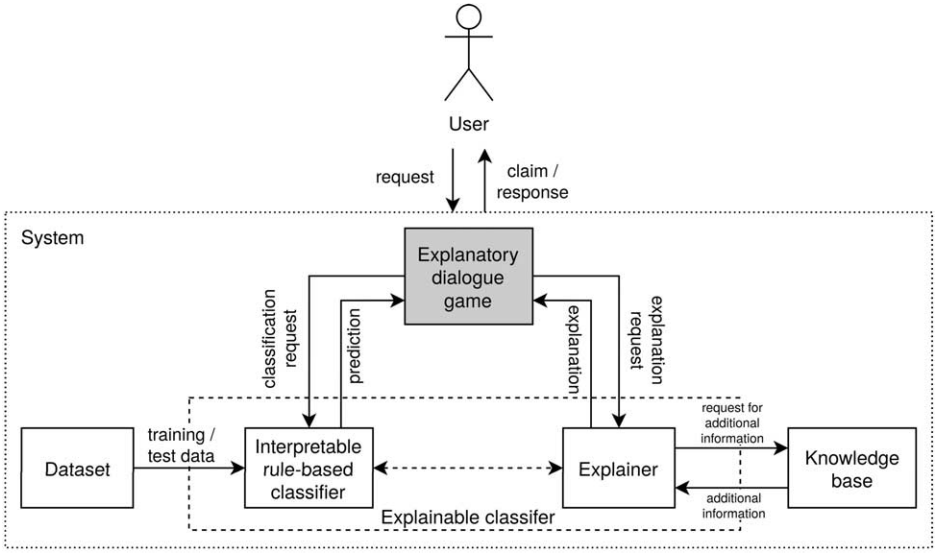


Fig. 1. A schema of the modelled system-user explanation communication process. This paper focuses on designing an explanatory dialogue game for communication of factual and counterfactual explanations for interpretable rule-based classifiers (the shaded block).

In this work, we assume knowledge of the feature space: the dataset is said to contain linearly scaled numerical features. In addition, all the numerical feature values are said to be mapped to the corresponding feature-dependent linguistic variable [86]. Therefore, each data instance $x_i \in X = \langle F_i, y_i \rangle$ is associated to class $y_i \in Y$ and defined over the set of p 3-tuple features $F_i = \{\langle f^k, v^k, t^k \rangle\}|_{k=1}^{p}$ where each feature f^k is assigned to the corresponding numerical value v^k and linguistic term t^k (e.g., ⟨age, 20, young⟩). The values of the linguistic variables (i.e., the so-called linguistic terms) may be defined by an expert. In this case, they are mapped to expert knowledge-based numerical intervals covering all the values of the corresponding feature. Otherwise, the linguistic variable is assigned to a set of textual values and mapped to equal-size numerical intervals. In this respect, the set of textual values that the linguistic variable can take on is of arbitrary cardinality.

The classifier predicts the class label $\hat{y}$ for the given test instance $x_{\text{test}} = \langle F_{\text{test}}, y_{\text{test}} \rangle$ on the basis of the learned mapping function c. The test instance classification is predicted correctly if the predicted class label and the actual test instance class label are the same (i.e., $\hat{y} = y_{\text{test}}$). Otherwise, the predicted class is deemed wrong (i.e., $\hat{y} \neq y_{\text{test}}$). Altogether, the interpretable rule-based classifier and the explainer are said to form an explainable classifier. Once the classifier outputs a prediction, the associated explainer attempts to generate an explanation in natural language for that prediction. Upon request, the explanation is passed to the User via the explanatory dialogue game, which serves as a communication channel between the explainable classifier and the User. During their intercourse, the User is assumed to be able to submit further explanation-related requests and receive responses processed by the dialogue game module whereas the dialogue game module can query the explainer for further explanation-related information.

2.2. *Explanation to the classification*

The upsurging need for explaining a classifier's output is raising interest in the mere nature of the explanation. For instance, social sciences testify that explanations are expected to be contrastive, selected,

and social [45]. First, the property of *contrastiveness* implies establishing a relation not only between the cause and effect of the phenomenon under consideration but also another relation between the cause and a given non-observed effect (i.e., another alternative effect). Second, explanations are as well argued to be *selected*, i.e., only the most relevant causes should make part of a specific explanation. Third, explanations are claimed to be *social*, i.e., they are a product of interaction between the explainer and the explainee.

Contrastiveness plays an important role when explaining a solution to the classification problem, as different classes are opposed to the others on the basis of distinctive feature values. Further, contrastiveness is inherent to counterfactual (CF) explanations (or counterfactuals, for short). In the context of XAI, counterfactuals suggest minimal changes in feature-value pairs for a different outcome to be obtained [71]. CFs are said to be post-hoc (i.e., they are generated for pretrained classifiers) and local (i.e., they explain the classifier's output w.r.t. a specific test instance) [27]. CFs may be (1) model-agnostic if they operate only on the given input (i.e., a test instance) and output (i.e., a prediction) of the classifier or (2) model-specific if they utilise the internals of the classifier to explain the given output [47,71].

CF explanations are claimed to have a number of desired properties against which CF explanation methods can be evaluated [27]. For example, CFs should be *valid* (i.e., CFs should truly lead to the desired hypothetical outcome), *proximate* (i.e., CFs should suggest only minimal changes to the test instance w.r.t. the selected distance metric), *sparse* (i.e., CFs should minimise the number of features whose values are to be changed), *actionable* (i.e., CFs should suggest feasible changes), and *diverse* (i.e., CFs should offer multiple alternatives). An exhaustive list of such properties can be found in recent surveys on CF explanation generation and evaluation [27,49,78]).

A large number of explanation generation methods are evaluated using automatically computable metrics that assess the aforementioned properties of CF explanations [49]. However, such metrics oftentimes do not take into consideration user feedback at all. Whereas considering the social factor may not be necessary when, e.g., measuring validity, estimating CF diversity may have to directly involve capturing effects of the interaction between the system and the user. Indeed, CF explanations suggesting minimal changes in feature values may not always be equally appreciated by end users. Given a variety of potential CFs, different users may prefer distinct CFs for the same hypothetical output. Further, the social aspect of explanation becomes crucially important when two alternative automatically generated pieces of explanation are deemed equally explanatory (e.g., when the distances from the test instance to two or more closest CF data points are the same or when two CF sets have the same coverage). As the state-of-the-art AI technologies are shifting towards being user-centric [83], it appears indispensable to enhance existing explanation generation modules with a system-user communication interface that would allow end users to produce such inquiries for alternative CFs in the course of an explanatory dialogue, even if the user is not aware of the dataset-related peculiarities.

Various state-of-the-art CF explanation generation frameworks are known to offer diverse CFs ([15,17,35,49,60,62,75,85], among others). However, the format of such CFs raises several important concerns. First, most of such frameworks lack any interaction with end users leaving the users without further guidance when interpreting the generated explanations. Second, some explainers output a set of distinct CFs altogether [49,60]. In these settings, the Grice's maxim of quantity [25] may be violated, as only a subset of the offered explanations can be sufficient for the end user. Third, a large number of diverse CF explanation generation frameworks provide their output in tabular form [15,17,35,49,62,75]. Whereas natural language generation tools can be used to transform tabular data into text, a taxonomy of necessary explanation-related requests and responses remains missing. To address these issues, we propose a transparent explanatory dialogue model for diverse factual and counterfactual explanation

communication that allows the end user to explore the explanation space iteratively until he or she can make an informed decision on whether the system's prediction can be trusted.

In light of the aforementioned considerations, a classifier's prediction can be explained factually and/or counterfactually. As we focus on the social factor of explanation generation in this paper, we assume that an explainer provides us with automatically generated textual factual and CF explanations operating in the settings described in Section 2.1. Below, we define both aforementioned types of explanation in terms of their linguistic realisation.

Driven by the assumptions above, both factual and CF explanations can be represented in two forms: using linguistic terms or numerical values (intervals). On the one hand, a purely textual explanation may be more intuitive and comprehensive to the explainee (e.g., "The test instance is of class Blanche because colour is pale and bitterness is low" or "The test instance would be of class Porter if colour were brown and strength were high"). On the other hand, explanations that incorporate numerical information may offer more detailed (and, perhaps, more precise) information while possibly requiring additional domain knowledge (e.g., "The test instance is of class Blanche because $0 \leqslant$ colour $\leqslant 3$ and $2 \leqslant$ bitterness $\leqslant 5$" or "The test instance would be of class Porter if colour ranged between 20 and 30 and strength ranged between 100 and 200"). In this work, we refer to explanations of both modalities as "high-level" and "low-level" explanations, respectively.

Definition 1. A *high-level explanation* $e^h(\hat{y}, [y'])$[1] is a set of feature-value pairs that explains the classifier's prediction $\hat{y}$ for the given test instance either factually or counterfactually in terms of the linguistic terms associated to the corresponding linguistic variable.

Definition 2. A *low-level explanation* $e^l(\hat{y}, [y'])$ is a set of feature-value pairs that explains the classifier's prediction $\hat{y}$ for the given test instance either factually or counterfactually in terms of the corresponding numerical values (intervals).

Paired explanations of both modalities may be found complementary to each other, as they may target different groups of end users. High-level explanations may facilitate understanding thereof by lay users. In turn, low-level explanations may be necessary for expert users to be able to further verify the validity of the offered explanation without linguistic ambiguity. Hereinafter, we assume that both factual and CF explanations to be paired two-level structures. To meet the requirement of being selective [45], all such explanations should be designed to reflect only the most characteristic features of the test instance that influence the classifier's prediction or its hypothetical counterpart. Let us now define factual and CF explanations in terms of their high- and low-level components.

Definition 3. A *factual explanation* $e_f(\hat{y}) = \langle e^h_f(\hat{y}), e^l_f(\hat{y})\rangle$ is a 2-tuple of affirmative sentences answering the question "Why is the test instance predicted to be of class $\hat{y}$?" where $e^h_f(\hat{y})$ and $e^l_f(\hat{y})$ are the corresponding high- and low-level explanations, respectively.

The given test instance's prediction can be explained in a (possibly, infinite) number of ways. At the same time, different explanations for the same phenomenon may have distinct degrees of explanatory power. Hence, all possible factual explanations are assumed to be ranked by an explainer in terms of their relevance to the test instance. Importantly, the notion of relevance in Definition 3 is determined by

[1] Hereinafter, $[y']$ is used as an optional parameter to refer to the requested CF class whenever a CF explanation is being processed. This parameter is omitted for the same request when a piece of factual explanation is being considered.

peculiarities of the explanation generation method, which falls outside the scope of this paper. The set of all factual explanations E_f for the predicted class $\hat{y}$ is defined as follows:

$$E_f(\hat{y}) = \bigcup_{i=1}^{\infty} e_{fi}(\hat{y}) \tag{1}$$

where e_{f1} is the most relevant factual explanation for the test instance's prediction, $e_{f|E_f|}$ is the least relevant one, i being the rank of the given piece of explanation. On the other hand, a CF explanation is assumed to suggest a minimal set of feature value changes that lead to a different desired classification. Then, a CF explanation is defined as follows.

Definition 4. A *counterfactual (or, shortly, CF) explanation* $e_{cf}(\hat{y}, y') = \langle e^h_{cf}(\hat{y}, y'), e^l_{cf}(\hat{y}, y')\rangle$ for the given CF class $y' \in Y \setminus \{\hat{y}\}$ is a 2-tuple of conditional sentences answering the question "Why is the test instance not predicted to be of class y' instead of $\hat{y}$?" where $e^h_{cf}(\hat{y}, y')$ and $e^l_{cf}(\hat{y}, y')$ are the corresponding high- and low-level explanations, respectively.

Similarly to factual explanations, all possible CFs are assumed to be ranked by their relevance to the test instance in accordance with a preselected criterion (for example, the distance metric from the test instance to the closest data point that the explanation includes). Then, the set of all the CF explanations for the given CF class is defined as follows:

$$E_{cf}\big(\hat{y}, y'\big) = \bigcup_{i=1}^{\infty} e_{cfi}\big(\hat{y}, y'\big) \tag{2}$$

where $e_{cf1}(\hat{y}, y')$ is the most relevant counterfactual explanation to the test instance's prediction $\hat{y}$ for the given CF class y', $e_{cf|E_{cf}|}$ is the least relevant one, i being the rank of an explanation.

Altogether, all ranked candidate factual and CF explanations for the given prediction are assumed to be unique and said to constitute an explanation space for the given prediction. The explanation space therefore contains all the pieces of factual and CF explanations that the system can offer to the end user w.r.t. the given test instance. Consequently, a given classifier's prediction cannot be explained by any piece of explanation that the explanation space does not contain.

Definition 5. An *explanation space* $E_{\text{space}}(\hat{y})$ is the union of all possible factual and CF explanations that an explainer can generate for the given prediction $\hat{y}$, s.t. $E_{\text{space}}(\hat{y}) = E_f(\hat{y}) \cup E_{cf}(\hat{y}, y'), \forall y' \in Y \setminus \{\hat{y}\}$.

2.3. Explainer-preferred vs. explainee-preferred explanations

Whereas any single piece of explanation may be satisfactory for the given user, it may have to be combined with other explanation instances for other users. For example, the end user may (1) request and be satisfied with the offered (factual and/or counterfactual) piece of explanation, (2) request and not be satisfied with the offered explanation, or (3) not request any explanation for, e.g. an alternative CF class, at all. In addition, not all the most relevant pieces of explanation from the system's point of view may seem as relevant to the user. To inspect the differences between such combinations of explanations, we therefore introduce the notions of explainer-preferred and explainee-preferred explanation. Explanation

rankings provided by the explainer allow us to single out the most relevant pieces of CFs for each CF class from the system's point of view:

$$E_{cf1}(\hat{y}) = \cup e_{cf1}\big(\hat{y}, y'\big), \quad \forall y' \in Y \setminus \{\hat{y}\} \tag{3}$$

Then, an explainer-preferred explanation is said to comprise all the most relevant (both factual and counterfactual) pieces of explanation from the explainer's point of view.

Definition 6. An *explainer-preferred explanation* is the union of the most relevant automatically generated factual explanation for the predicted class and the most relevant explanations for each of the CF classes:

$$E_{\text{explainer}}(x_{\text{test}}, \hat{y}) = e_{f1}(\hat{y}) \cup E_{cf1}(\hat{y}) \tag{4}$$

An explainer-preferred explanation may be claimed to comprehensively explain the output of the given classifier to any end user. Given a set of multiple candidate factual and/or counterfactual explanations from the explanation space, the explanation generation module ranks them by relevance to the test instance (e.g., a distance metric) and subsequently presents the most relevant pieces of explanation to the end user. However, the explanation generation module output ignores end user preferences in these settings. Therefore, we define an explainee-preferred explanation as follows.

Definition 7. An *explainee-preferred explanation* is the union of all the pieces of explanation that the explainee finds the most satisfactory, as he or she explores the explanation space E_{space} when being explained the given prediction.

For illustrative purposes, consider the classification task for a dataset of four classes: $Y = \{y_1, y_2, y_3, y_4\}$. Let some test instance x_{test} be predicted to be of class y_1. An explainer-preferred explanation would therefore include the most relevant piece of factual explanation for class y_1 as well as the most relevant explanations for all the other (CF) classes:

$$E_{\text{explainer}}(x_{\text{test}}, y_1) = e_{f1}(y_1) \cup e_{cf1}(y_1, y_2) \cup e_{cf1}(y_1, y_3) \cup e_{cf1}(y_1, y_4) \tag{5}$$

The explainee may consider (a part of) the offered explanation irrelevant, redundant, or poorly explanatory. Figure 2 illustrates a possible discrepancy between the automatically generated and some user-preferred explanations. Whereas the factual explanation may be satisfactory for him or her, the explainee may find optimal the third most relevant CF explanation (from the explainer's point of view) for class y_2 (if it were offered), the second most relevant CF explanation for class y_3, and not require any CF explanation for class y_4. In this case, the reconstructed user-preferred explanation could be formally represented as follows:

$$E_{\text{explainee}}(x_{\text{test}}, y_1) = e_{f1}(y_1) \cup e_{cf3}(y_1, y_2) \cup e_{cf2}(y_1, y_3) \tag{6}$$

As shown in the example above, there may exist only a slight overlap between the most relevant explainer-preferred explanation and that expected by the explainee. It therefore appears indispensable to provide end users with a means of interaction with the explanation generation module to enable them to interactively explore the explanation space and, subsequently, shape the explanation in accordance

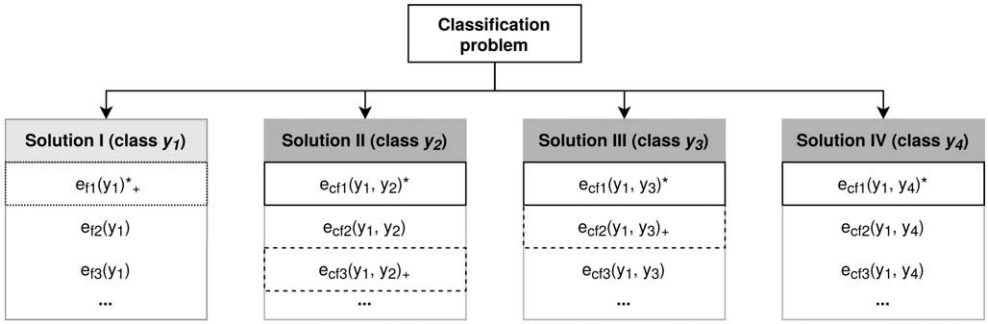


Fig. 2. A schema of a classification problem. Class y_1 is predicted by the classifier to be the solution to the problem. The other possible solutions (classes y_2, y_3, y_4) are considered hypothetical and form the set of CF classes. The corresponding explanations in solid rectangles (additionally marked with * as superscript) are those generated automatically. The explanations in dashed rectangles (additionally marked with + as subscript) are those preferred by the end user. Notably, the factual explanation in a double-dashed rectangle (that for class y_1) is both explainer-preferred and explainee-preferred.

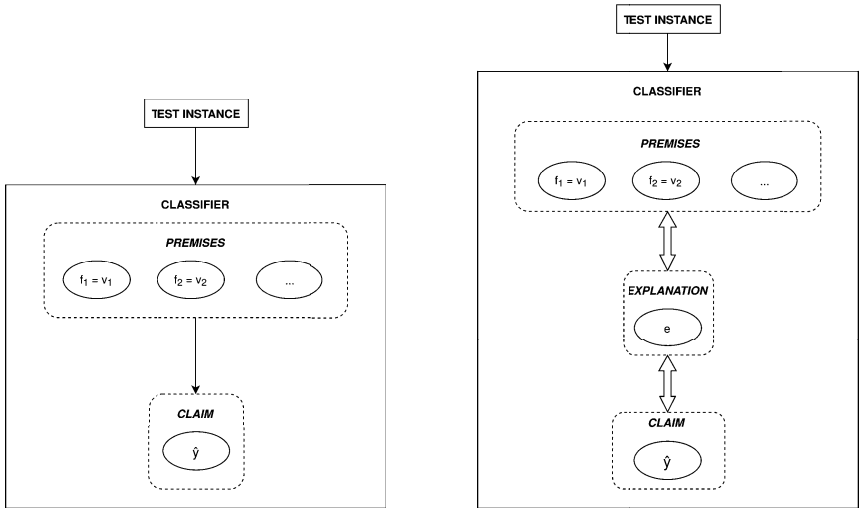
with their preferences. To do so, it is helpful to consider the classifier's reasoning from the argumentative point of view. Argumentation is regarded as an effective mechanism to communicate explanation in natural language [8]. Thus, various argumentation frameworks are shown to be particularly useful in the field of XAI for their ability to generate explanations of different modalities (e.g., textual, graphical, hybrid) [16]. Further, recent work on argumentation-based explanation generation shows that such frameworks provide efficient explanatory interfaces between AI-based systems and users of such systems, particularly, in the form of dialogue [77]. In addition, argumentation is shown to logically connect with, for example, abductive reasoning tools that are widely used for counterfactual reasoning [11].

In these settings, a prediction may be treated as a claim proposed by the classifier. Such a claim is then supported by the decisive feature value pairs (either specific values or intervals of values) that led the classifier to make the corresponding prediction (see Fig. 3(a)). However, ground-truth data-based premises cannot be attacked directly, as they can by no means be claimed invalid. Therefore, it appears necessary to introduce an intermediate explanation layer that approximates the premises and serves as an attackable natural language interface between the premise and the claim (see Fig. 3(b)).

Throughout this paper, we claim that rule-based explanations from interpretable classifiers serve this purpose well. First, they reflect the features retrieved from the data that the classifier is trained on. Second, their natural language representation allows the end user to construct a comprehensive mental representation of the underlying data. Following Hempel's definition of explanation [31], explanations themselves can be regarded as arguments. In the context of explanatory dialogue between the system and the user, explanations can then be attacked in the dialogic intercourse between the dialogue parties. In this manner, the end user is given the opportunity to interactively inspect explanations from the explanation space that do not make part of the explainer-preferred explanation by arguing over the initially (and, if necessary, also subsequently) offered pieces thereof.

3. Dialogue game for XAI

In this section, we formally define a dialogue game that serves to communicate explanation(-s) generated automatically by an explanation generation module (paired with the corresponding interpretable rule-based classifier) to its end user. Thus, Section 3.1 proposes formal components of explanatory di-



(a) A premise-claim representation of a classifier's prediction.

(b) A premise-explanation-claim representation of a classifier's prediction.

Fig. 3. Schematic representations of classifier's reasoning from the argumentative point of view.

alogue. Subsequently, Section 3.2 presents an example of an explanatory dialogue modelled in accordance with the principles outlined in Section 3.1. Finally, Section 3.3 generalises the proposed approach to explanatory information-seeking dialogue modelling in form of an explanatory context-free dialogue grammar.

3.1. Formal description of explanatory information-seeking dialogue

In order to construct a communication channel between the system and the end user, we propose that explanatory dialogue be modelled on the basis of the so-called "dialogue game" approach to argumentation [54]. Taking into consideration the aforementioned requirements to explanation, we formally define an explanatory dialogue between the explanation generation module and end user as a 10-tuple $D = \langle P, M, R, Pr, K, E, DET, CLAR, CFS, KB \rangle$ where

- P is the set of dialogue participants;
- M is the set of dialogue moves that the dialogue participants make in the course of a dialogue;
- R is the set of requests and responses that specify allowed utterances in the course of explanatory dialogue;
- Pr is the dialogue protocol governing the flow of the conversation in accordance with the set of predetermined locution rules specifying types of legitimate utterances;
- K is the knowledge store, i.e. the dynamically populated set of all the pieces of explanation that the user requests and receives during his or her interaction with the system;
- E is the explanation store, i.e. the dynamically updated set of the last offered pieces of explanation for each class under consideration;
- DET is the detailisation store, i.e. the set of features of the actually processed piece of high-level explanation whose values (i.e., linguistic terms) can be inspected for further details;
- $CLAR$ is the clarification store, i.e. the set of features of the actually processed piece of explanation whose definitions can be requested;

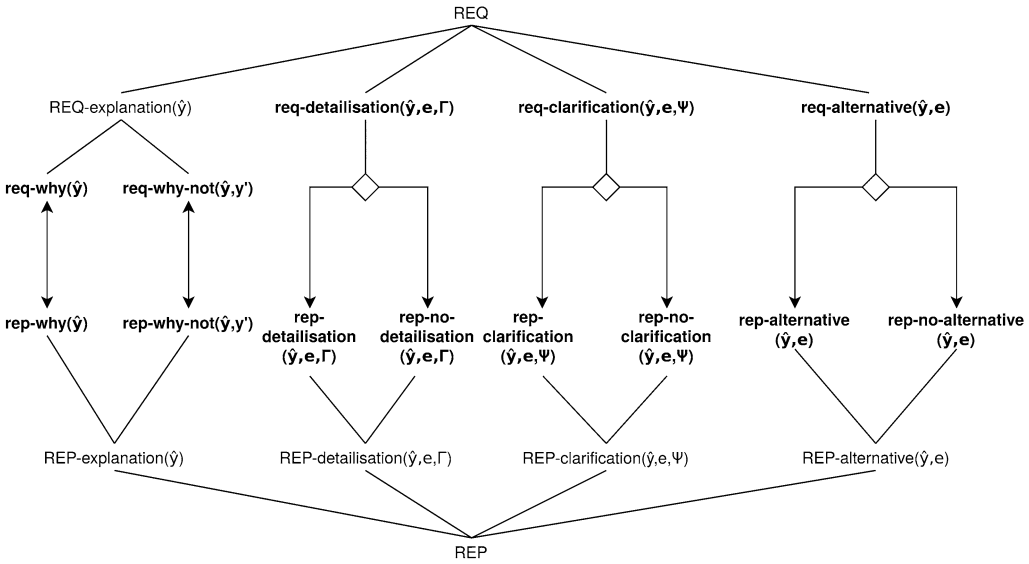


Fig. 4. A typology of requests and replies. Individual requests/responses are in bold. In addition, sets of request/responses are named with uppercase letters (i.e., REQ-/REP-).

- *CFS* is the CF class store, i.e. the set of CF classes whose explanations can be potentially offered to the end user;
- *KB* is a knowledge base containing the domain knowledge for the addressed problem.

Let us now define each component of the proposed explanatory dialogue model in detail.

1) Participants. An explanatory dialogue serves as an interface between two parties: the explainable classifier (or, in general, the system S) and the human agent interacting with the system (the user U). Therefore, the set of participants is defined to always consist of two items $P = \{S, U\}$ where the system S always plays the role of the explainer whereas the user U is always the explainee.

2) Moves. A single instance of a dialogue can be regarded as a sequence of finite legitimate moves $M = \langle m_0, m_1, \ldots, m_n \rangle$, each of which is generated in accordance with the locution rules as well as those making part of the corresponding dialogue protocol.

3) Responses and requests. Our explanatory dialogue model presupposes that the explainer (i.e., the system) has the ability to present all the information available to it to the explainee (i.e., the user). The user is, in turn, capable of inquiring all such information. It is therefore crucially important to find a balance between the information that the user may require from the system and the information the system can provide the user with.

Driven by the assumption that high- and low-level explanations may accommodate both expert and lay users and inspired by previous work on formal explanatory dialogue modelling [9], we distinguish four types of user requests and responses that form the corresponding set $R = \{REQ, REP\}$. Namely, those are the requests for (either factual or CF) explanation, detailisation, clarification, and alternative explanation of either of the considered kinds. Figure 4 summarises all possible types of user's requests and the corresponding system's responses. All locutions generated by both parties fall into either of the two symmetric classes.

On the one hand, the set of requests from the user to the system *REQ*={*REQ-explanation*($\hat{y}$), *req-detailisation*($\hat{y}$, e, Γ), *req-clarification*($\hat{y}$, e, Ψ), *req-alternative*($\hat{y}$, e)} consists of the following items:[2]

- *REQ-explanation*($\hat{y}$): the set of user requests for explanation for system's prediction $\hat{y}$;
- *req-detailisation*($\hat{y}$, e, Γ): the user request for further details on feature Γ (i.e., the corresponding numerical intervals) that makes part of a high-level (either factual or CF) explanation e for prediction $\hat{y}$;
- *req-clarification*($\hat{y}$, e, Ψ): the user request for clarification of the meaning of a specific feature Ψ that makes part of (either factual of CF and either high-level or low-level) explanation e for prediction $\hat{y}$;
- *req-alternative*($\hat{y}$, e): the user request for an alternative (either factual or CF and either high-level or low-level) explanation provided that the user is not satisfied with the previously offered explanation e for system's prediction $\hat{y}$.

Further, the set of user explanation requests *REQ-explanation*($\hat{y}$) consists of the following possible locutions:

- *req-why*($\hat{y}$): the user request for a factual explanation for the system's prediction $\hat{y}$;
- *req-why-not*($\hat{y}$, y'): the user request for a CF explanation concerning the CF class $y' \in Y \setminus \{\hat{y}\}$ for prediction $\hat{y}$ (i.e., to specify why some CF class y' was not predicted instead of $\hat{y}$).

On the other hand, the set of responses (replies) that the system sends back to the user *REP*={*REP-explanation*($\hat{y}$), *REP-detailisation*($\hat{y}$, e, Γ), *REP-clarification*($\hat{y}$, e, Ψ), *REP-alternative*($\hat{y}$, e)} mirrors the set of user requests:

- *REP-explanation*($\hat{y}$): the set of system responses in an attempt to explain prediction $\hat{y}$;
- *REP-detailisation*($\hat{y}$, e, Γ): the set of system responses in an attempt to provide details (i.e., numerical intervals) with respect to feature Γ of explanation e for system's prediction $\hat{y}$;
- *REP-clarification*($\hat{y}$, e, Ψ): the set of system responses in an attempt to clarify feature Ψ making part of (either factual or CF) explanation e for prediction $\hat{y}$;
- *REP-alternative*($\hat{y}$, e): the set of system responses in an attempt to provide the user with an explanation alternative to the previously offered (either factual or CF and either high-level or low-level) explanation e for prediction $\hat{y}$.

In addition, the set of replies to requests for (initial, non-alternative) explanation *REP-explanation*($\hat{y}$) consists of the following items:

- *rep-why*($\hat{y}$): the system attempts to factually explain the prediction $\hat{y}$ on the basis of the known features that led to that decision and offers a factual explanation if it is able to, or refuses to offer it, otherwise;
- *rep-why-not*($\hat{y}$, y'): the system attempts to provide the user with a CF explanation for prediction $\hat{y}$ for the given CF class y' or refuses to offer it, otherwise.

The set of replies to detailisation requests *REP-detailisation*($\hat{y}$, e, Γ) consists of the following items:

- *rep-detailisation*($\hat{y}$, e, Γ): the system provides the numerical intervals over which the corresponding linguistic term of the requested explanation feature Γ making part of explanation e is defined;

[2]Sets of requests are denoted using uppercase letters (as in, e.g., *REQ-explanation*) whereas single instances of requests are denoted using only lowercase letters (as in, e.g., *req-detailisation*).

- *rep-no-detailisation*($\hat{y}$, e, Γ): the system refuses to provide numerical intervals on the requested feature's linguistic term in explanation e, e.g. due to their unavailability.

The set of replies to clarification requests *REP-clarification*($\hat{y}$, e, Ψ) consists of the following items:

- *rep-clarification*($\hat{y}$, e, Ψ): the system provides the user with a definition of the requested feature Ψ making part of explanation e for prediction $\hat{y}$ retrieving it from the knowledge base;
- *rep-no-clarification*($\hat{y}$, e, Ψ): the system refuses to clarify the requested feature Ψ making part of explanation e for prediction $\hat{y}$ due to, e.g., its absence in the knowledge base.

The set of replies to alternative explanation requests *REP-alternative*($\hat{y}$, e) consists of the following items:

- *rep-alternative*($\hat{y}$, e): the system recognises the fact that the user is not satisfied with the offered (factual or CF) explanation e for prediction $\hat{y}$, seeks the most relevant alternative to it, generates and offers an alternative explanation to the user;
- *rep-no-alternative*($\hat{y}$, e): the system recognises the fact that the user is not satisfied with the offered (factual or CF) explanation e for prediction $\hat{y}$, seeks the most relevant alternative to it, but is unable to generate it.

4) Dialogue protocol. An explanatory dialogue between the system and the user is modelled following the rules specified in the dialogue protocol. The protocol determines turntaking rules, the rules governing user's and system's allowed moves at each stage of the explanatory dialogue, and the termination states of the dialogue. Thus, the locution types above are directly mapped to the speech acts produced by the system and the user as specified in the dialogue protocol. All of the aforementioned protocol rules are specified in Appendix B.

5) Knowledge store. Let K be the knowledge store which accumulates user's knowledge w.r.t. explanations requested during his or her interaction with the system. Knowledge store K is initialised to be an empty set: $K = \emptyset$. When the system generates a factual or CF explanation (locutions *explain-f*($\hat{y}$, E, e_f) and *explain-cf*($\hat{y}$, E, y', e_{cf}), as specified in the dialogue protocol), the corresponding piece of explanation is added to the knowledge store: $K = K \cup e_f(\hat{y})$ or $K = K \cup e_{cf}(\hat{y}, y')$, respectively. The same applies to alternative explanations of either kind (locutions *alter-f*($\hat{y}$, E, e_f, e'_f) and *alter-cf*($\hat{y}$, E, y', e_{cf}, e'_{cf})).

6) Explanation store. Let E be the explanation store which tracks the current state of the explainee-preferred explanation throughout the dialogue. Explanation store E is initialised to be an empty set: $E = \emptyset$. Similarly to the knowledge store, a factual or CF explanation is added to the explanation store once generated: $E = E \cup e_f(\hat{y})$ or $E = E \cup e_{cf}(\hat{y}, y')$, respectively. If the user finds the offered factual or CF explanation not satisfactory enough and asks for an alternative explanation (locutions *why-alternative*($\hat{y}$, E, e_f) and *why-not-alternative*($\hat{y}$, E, y', e_{cf}), respectively), the corresponding explanation is removed from the explanation store: $E = E \setminus e_f(\hat{y})$ or $E = E \setminus e_{cf}(\hat{y}, y')$, respectively. Noteworthy, the user cannot request an alternative explanation to any explanation non-offered previously. Further, the user can only submit explanation-related requests (detailisation, clarification, alternative) for the piece of explanation being processed. The resulting explainee-preferred explanation is the union of all the pieces of explanation found in the explanation store when a terminal dialogue state is reached.

7) Detailisation store. Let *DET* be the store that contains the features of the currently processed high-level explanation for which further details can be requested. *DET* is initialised to be empty, as the explanatory dialogue starts: $DET = \emptyset$. The user can submit a detailisation request to the system only if a high-level (either factual or CF) explanation $e = e^h_f | e^h_{cf}$ is being processed. Recall that for

each feature Γ of the currently processed high-level explanation e, the feature is defined in terms of a linguistic variable mapped to the corresponding linguistic terms. When a new piece of high-level explanation is offered to the end user, *DET* is reinitialised with the set of features that the currently processed explanation contains: $DET = \{\Gamma\}, \forall\Gamma \in e$. The user can ask the system to provide him or her with the numerical intervals for the linguistic term of the given explanation feature only once during a sub-dialogue concerning a specific piece of explanation. Thus, the corresponding feature is eliminated from the detailisation store once the system has generated a response θ (locution *elaborate*($\hat{y}$, $E[,y']$, e, Γ, θ)): $DET = DET \setminus\{\Gamma\}$. If $DET = \emptyset$, it is prohibited for the user to submit a detailisation request (locution *what-details*($\hat{y}$, $E[,y']$, e, Γ)). When the user makes the final decision w.r.t. the system's claim (i.e., either accepts or rejects it), the detailisation store is nullified: $DET = \emptyset$.

8) Clarification store. Let *CLAR* be the clarification store that contains the explanation features whose meaning can be clarified. Similarly to the detailisation store, *CLAR* is initialised to be empty: $CLAR = \emptyset$. When a new piece of explanation is offered, *CLAR* is populated with all the features that the explanation being processed $e = e^h_f|e^h_{cf}|e^l_f|e^l_{cf}$ contains: $CLAR = \{\Psi\}, \forall\Psi \in e$. Noteworthy, the definitions for all the features that the dataset contains are precollected, mapped to one another by an expert or retrieved from a dictionary, and stored in the knowledge base. The user can ask to clarify a specific feature from the clarification store only once during a sub-dialogue concerning a specific piece of explanation. Then, the corresponding feature is eliminated from the clarification store after the system's response υ (locution *clarify*($\hat{y}$, $E[,y']$, e, Ψ, υ)): $CLAR = CLAR \setminus\{\Psi\}$. If $CLAR = \emptyset$, it is prohibited for the end user to submit a clarification request (locution *what-is*($\hat{y}$, $E[,y']$, e, Ψ)). When the user makes the final decision w.r.t. the system's claim (i.e., either accepts or rejects it), the clarification store is nullified: $CLAR = \emptyset$.

9) CF class store. Let *CFS* be the CF class store that contains all CF classes. It is initialised upon the successful execution of the factual explanation request (locution *explain-f*($\hat{y}$, E, e_f)) so that $CFS = Y \setminus \{\hat{y}\}$ for some prediction $\hat{y} \in Y$. The user is allowed to request a CF explanation for each class from *CFS* only once (locution *why-not-explain*($\hat{y}$, E, y')). In addition, the user is allowed to ask for a (series of) alternative CF explanation(-s) for the same CF class (locution *why-not-alternative*($\hat{y}$, E, y', e_{cf}) as many times as there are alternative CFs for that class. Once a CF explanation is requested for some CF class y', it is eliminated from the *CFS* store: $CFS = CFS \setminus\{y'\}$. When the user makes the final decision w.r.t. the system's claim (i.e., either accepts or rejects it), the CF class store is nullified: $CFS = \emptyset$.

10) Knowledge Base. The knowledge base contains the dataset-related domain knowledge including a specification of all the dataset features (e.g., linguistic terms, the corresponding intervals, and definitions of all the features that the dataset contains).

3.2. Illustrative example

Having introduced the proposed formalism for explanatory information-seeking dialogue modelling, let us now illustrate it taking the previously considered example for reference (see Table 1 for details). Thus, we are considering the beer style classification problem for the beer dataset that contains the following classes: $Y_{\text{beer}} = \{$*Blanche, Lager, Pilsner, IPA, Barleywine, Stout, Porter, Belgian strong ale*$\}$. Table 2 outlines the states of the detailisation, clarification, and CF class stores of the example explanatory dialogue after each dialogue move. Table 3 outlines the states of the knowledge and explanation stores for the same example dialogue.

Initially, the system claims that some instance of beer is of class *Blanche* (move m_1). All the stores that make part of the dialogue model (*K, E, DET, CLAR, CFS*) are initialised to be empty. At the next

Table 2
A move-by-move formal description of the stores governing the example of explanatory dialogue from Table 1

Move	*Locution*	*DET*	*CLAR*	*CFS*
m_1	*claim* $(\hat{y}, E)$	$\emptyset$	$\emptyset$	$\emptyset$
m_2	*why-explain* $(\hat{y}, E)$	$\emptyset$	$\emptyset$	$\emptyset$
m_3	*explain-f* $(\hat{y}, E, e_f)$	{*colour, bitterness*}	{*colour, bitterness*}	{*Lager, Pilsner, IPA, Barleywine, Stout, Porter, Belgian strong ale*}
m_4	*what-is*$(\hat{y}, E, e_f, \Psi)$	{*colour, bitterness*}	{*colour, bitterness*}	{*Lager, Pilsner, IPA, Barleywine, Stout, Porter, Belgian strong ale*}
m_5	*clarify* $(\hat{y}, E, e_f, \Psi, \upsilon)$	{*colour, bitterness*}	{*colour*}	{*Lager, Pilsner, IPA, Barleywine, Stout, Porter, Belgian strong ale*}
m_6	*why-not-explain* $(\hat{y}, E, y')$	{*colour, bitterness*}	{*colour*}	{*Lager, Pilsner, IPA, Barleywine, Stout, Porter, Belgian strong ale*}
m_7	*explain-cf* $(\hat{y}, E, y', e_{cf})$	{*colour, bitterness*}	{*colour, bitterness*}	{*Lager, Pilsner, IPA, Barleywine, Porter, Belgian strong ale*}
m_8	*what-details* $(\hat{y}, E, e_{cf}, \Gamma)$	{*colour, bitterness*}	{*colour, bitterness*}	{*Lager, Pilsner, IPA, Barleywine, Porter, Belgian strong ale*}
m_9	*elaborate* $(\hat{y}, E, e_{cf}, \Gamma, \theta)$	{*colour*}	{*colour, bitterness*}	{*Lager, Pilsner, IPA, Barleywine, Porter, Belgian strong ale*}
m_{10}	*why-not-explain*$(\hat{y}, E, y'')$	{*colour*}	{*colour, bitterness*}	{*Lager, Pilsner, IPA, Barleywine, Porter, Belgian strong ale*}
m_{11}	*explain-cf*$(\hat{y}, E, y'', e_{cf})$	{*colour*}	{*colour*}	{*Lager, Pilsner, IPA, Barleywine, Belgian strong ale*}
m_{12}	*why-not-alternative*$(\hat{y}, E, y'', e_{cf})$	{*colour*}	{*colour*}	{*Lager, Pilsner, IPA, Barleywine, Belgian strong ale*}
m_{13}	*alter-cf*$(\hat{y}, E, y'', e_{cf}, e'_{cf})$	{*colour, strength*}	{*colour, strength*}	{*Lager, Pilsner, IPA, Barleywine, Belgian strong ale*}
m_{14}	*accept-u* $(\hat{y}, E)$	$\emptyset$	$\emptyset$	$\emptyset$
m_{15}	*accept-s* $(\hat{y}, E)$	$\emptyset$	$\emptyset$	$\emptyset$

step, the user requests a factual explanation for the given prediction (m_2). The system provides the user with a factual explanation (m_3). As the factual explanation is generated, both *DET* and *CLAR* stores are populated with the corresponding features (colour and bitterness). Further, the piece of factual explanation $e_f(\hat{y} = \text{Blanche})$ is placed to both the knowledge store and the explanation store. In addition, the CF store *CFS* is populated with all the CF classes. At the next stage, the user asks the system to clarify the notion of bitterness (m_4) and receives the corresponding definition from the system (m_5). As the clarification request for a given feature can only be submitted once while processing a specific piece of explanation, *bitterness* is then eliminated from the *CLAR* store.

Once the factual explanation is offered, the user may commit to the factual explanation offered and inquire a CF explanation for some CF class. In the present example, the user seeks, at this stage, to know why the classifier did not predict the given beer to be *Stout* (m_6). Then, the classifier presents the most relevant piece of CF explanation for this CF class in accordance with its ranking (m_7). The CF explanation $e_{cf}(y' = \text{Stout})$ is then added to both the knowledge and explanation stores, whereas the class *Stout* is removed from the *CFS* store. Then, the *DET* and *CLAR* stores are updated with the features that the newly offered CF explanation contains. As the user requires more detailed information on *bitterness* (m_8), the system retrieves the requested numerical interval over which the value of *bitterness* is defined to be high (m_9). The feature *bitterness* is then removed from the *DET* store. Then, the user proceeds to request a CF explanation for class *Porter* (m_{10}). Similarly to the previously offered explanations, *DET* and *CLAR* are updated accordingly, as the most relevant piece (from explainer's point of view) of CF

Table 3
An example explanatory dialogue schema

Block	*Move*	*Utterance*	K	E
C	m_1	**System:** The test instance is of class y.	$\emptyset$	$\emptyset$
E	m_2	**User:** Could you explain why you think so?	$\emptyset$	$\emptyset$
	m_3	**System:** It is of class y because $\langle \text{feature}_1 \rangle$ is $\langle \text{term}_1 \rangle$.	$\{e_f(\hat{y})\}$	$\{e_f(\hat{y})\}$
	m_4	**User:** What do you mean by $\langle \text{feature}_1 \rangle$?	$\{e_f(\hat{y})\}$	$\{e_f(\hat{y})\}$
	m_5	**System:** $\langle \text{feature}_1 \rangle$ is $\langle \text{definition } for \text{ feature}_1 \rangle$.	$\{e_f(\hat{y})\}$	$\{e_f(\hat{y})\}$
	m_6	**User:** But why is it not of class y'?	$\{e_f(\hat{y})\}$	$\{e_f(\hat{y})\}$
	m_7	**System:** It would be of class y' if $\langle \text{feature}_1 \rangle$ were $\langle \text{term}_2 \rangle$ and $\langle \text{feature}_2 \rangle$ were $\langle \text{term}_3 \rangle$.	$\{e_f(\hat{y}), e_{cf}(y')\}$	$\{e_f(\hat{y}), e_{cf}(y')\}$
	m_8	**User:** Could you specify how $\langle \text{feature}_1 \rangle$ is defined?	$\{e_f(\hat{y}), e_{cf}(y')\}$	$\{e_f(\hat{y}), e_{cf}(y')\}$
	m_9	**System:** $\langle \text{feature}_1 \rangle$ is defined to be $\langle \text{term}_2 \rangle$ because it is found in the interval $\langle [\text{term}_{2\text{min}}, \text{term}_{2\text{max}}] \rangle$.	$\{e_f(\hat{y}), e_{cf}(y')\}$	$\{e_f(\hat{y}), e_{cf}(y')\}$
	m_{10}	**User:** But why is the test instance not of class y''?	$\{e_f(\hat{y}), e_{cf}(y')\}$	$\{e_f(\hat{y}), e_{cf}(y')\}$
	m_{11}	**System:** It would be of class y'' if $\langle \text{feature}_1 \rangle$ were $\langle \text{term}_3 \rangle$ and $\langle \text{feature}_3 \rangle$ were $\langle \text{term}_3 \rangle$.	$\{e_f(\hat{y}), e_{cf}(y'), e_{cf}(y'')\}$	$\{e_f(\hat{y}), e_{cf}(y'), e_{cf}(y'')\}$
	m_{12}	**User:** I am not quite satisfied with your explanation. Could you offer me another one?	$\{e_f(\hat{y}), e_{cf}(y'), e_{cf}(y'')\}$	$\{e_f(\hat{y}), e_{cf}(y')\}$
	m_{13}	**System:** Sure! It would be of class y'' if...	$\{e_f(\hat{y}), e_{cf}(y'), e_{cf}(y''), e'_{cf}(y'')\}$	$\{e_f(\hat{y}), e_{cf}(y'), e'_{cf}(y'')\}$
T	m_{14}	**User:** Okay, I trust your prediction.	$\{e_f(\hat{y}), e_{cf}(y'), e_{cf}(y''), e'_{cf}(y'')\}$	$\{e_f(\hat{y}), e_{cf}(y'), e'_{cf}(y'')\}$
	m_{15}	**System:** Thank you for your trust in me. Bye!	$\{e_f(\hat{y}), e_{cf}(y'), e_{cf}(y''), e'_{cf}(y'')\}$	$\{e_f(\hat{y}), e_{cf}(y'), e'_{cf}(y'')\}$

In the left-hand side column ("Block"), C stands for claim, E – for explanation, T – for termination).

explanation is generated and offered for the class *Porter* (m_{11}). Then, the class *Porter* is excluded from the *CFS* store whereas the newly offered CF explanation is added to the knowledge and explanation stores. However, as the user is left dissatisfied or not convinced enough with the offered explanation, he or she inquires an alternative explanation to the previously offered one (m_{12}). Then, the latest offered explanation is removed from the explanation store. Subsequently, if the next best ranked alternative can be offered, it is added to the explanation store (m_{13}). The *DET* and *CLAR* stores are then updated accordingly. Having processed the presented explanations in their entirety, the user makes an informed decision that the classifier's prediction can be accepted (m_{14}). The system terminates the dialogue outputting a farewell locution (m_{15}).

Table 3 generalises the presented example of explanatory dialogue for any dataset where features, linguistic terms, and classes serve as dataset-specific variables. It is possible to generalise any explanatory dialogue modelled in accordance with the proposed framework using the suggested template utterances. Noteworthy, three main building blocks of such explanatory dialogue (C – claim, E – explanation, and T – termination) can be distinguished. Figure 5 presents the corresponding (partial, for illustrative purposes) parse tree of such a generalised explanatory dialogue.

3.3. Explanatory dialogue grammar (EDG)

As follows from the example of dialogue presented in Section 3.2, the proposed dialogue model has a hierarchical structure with respect to its main building blocks. This observation allows us to reflect the modular composition of explanatory dialogue (following our model) in a context-free dialogue grammar. As the transitions between the states of the dialogue are finite and predefined, the use of the correspond-

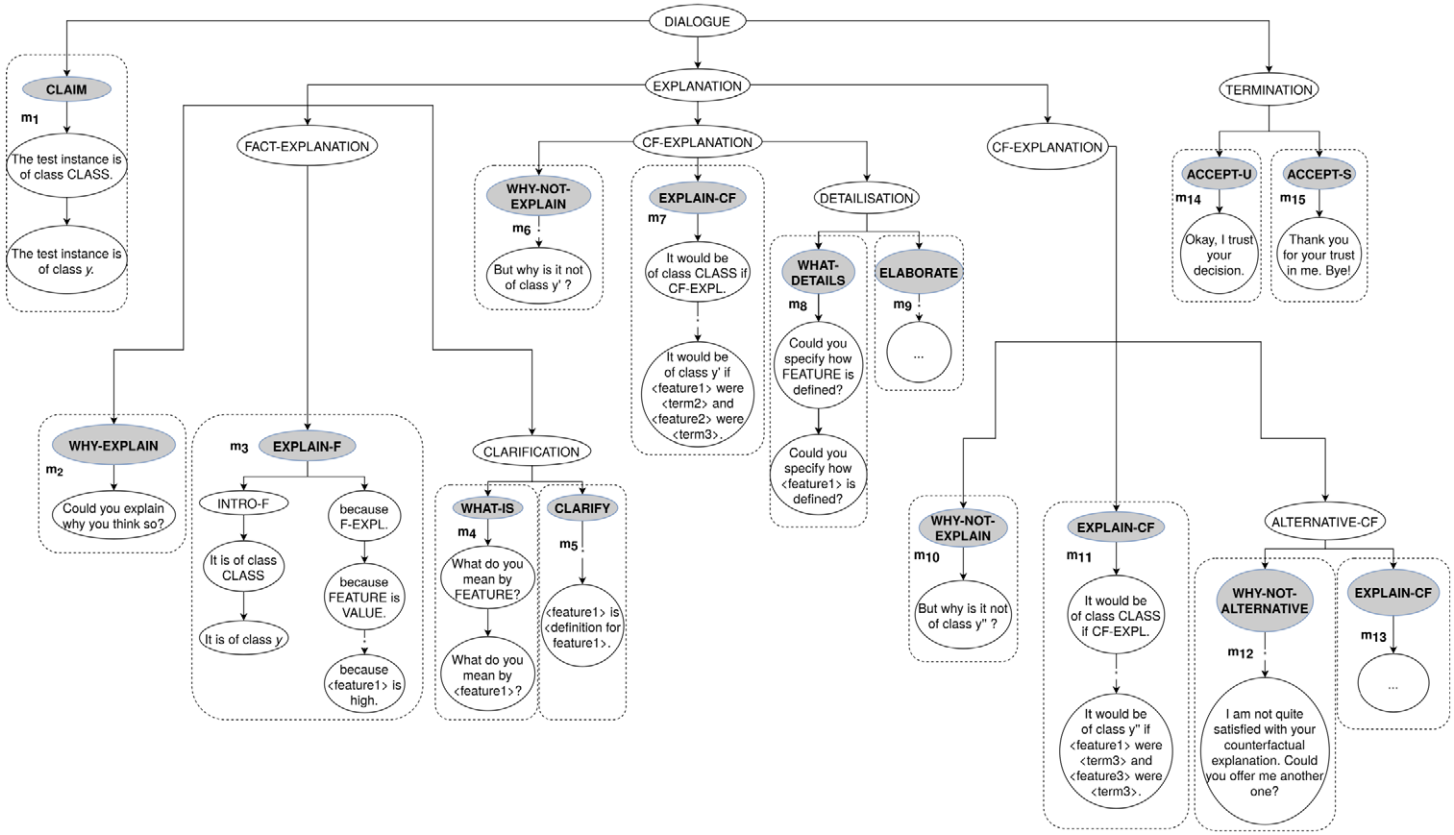


Fig. 5. A parse tree of the example of explanatory dialogue. Shaded nodes are non-terminals corresponding to specific speech acts. The subtrees in the dashed regions represent dialogue moves.

ing EDG allows us to (1) generate any explanatory dialogue that is valid in accordance with the dialogue protocol restrictions and (2) parse any actually valid explanatory dialogue or make a conclusion that the present explanatory dialogue is invalid with respect to the dialogue model constraints. Further, a grammar-based dialogue model can take into account modifications in the dialogue protocol if those are deemed necessary.

In light of the above, we define an EDG following Chomsky's definition of a context-free grammar as a tuple $G = \langle T, N, P, S \rangle$ where T is the set of terminals, N is the set of non-terminals, P is the set of production rules (productions), and S is the start token. In our model, T corresponds to a sentence actually uttered by each participant in the course of a dialogue. N encompasses the internal building blocks of the dialogue as well as the speech acts involved (see the shaded nodes in Fig. 5 for details). Thus, any explanatory dialogue is said to have three main building blocks (those corresponding to the non-terminals *CLAIM, EXPLANATION, TERMINATION*). In accordance with current legal requirements to explanation for AI, the block *EXPLANATION* enables the user to exercise the right to explanation and is made optional. All the non-terminals produced from the non-terminal *EXPLANATION* are designed in accordance with the predefined requests and responses (see Section 3.1 for details). In addition, P is composed in accordance with the dialogue protocol settings (see Appendix B for details). Note that productions can be subdivided in two groups, i.e., dataset-independent and dataset-specific productions. Dataset-independent production rules form the core of the proposed explanatory dialogue model and can be used in any application domain so long as it meets the settings of the classification problem as described in Section 2.1. The dataset-independent rules valid for the illustrative example of an explanatory dialogue are outlined in Appendix C. In turn, dataset-specific rules follow the structure of the given dataset and they are restricted by the information provided by the given interpretable rule-based classifier and the corresponding knowledge base. Finally, the start token S is known to always be the non-terminal *DIALOGUE* node, i.e., the root node in the tree depicted in Fig. 5.

4. Process mining for dialogue analytics

The proposed model of explanatory dialogue is designed in a top-down manner, which signals certain shortcomings. Thus, the dialogue protocol bases on the assumption that the taxonomy of requests and responses proposed in Section 3 inspired by findings from the literature exhaustively covers user's needs and system's abilities when engaged in an explanatory dialogue. However, in the absence of any empirical evaluation, such assumptions may result being purely speculative. For example, specific requests may be utilised to a very limited extent or even not utilised at all. Alternatively, there may exist requests that are not included in the original model, which may nevertheless be considered essential for human-machine interaction by the explainees. Either way, modifications to the model should be grounded on the data obtained from the end users. As such data-driven conclusions on the utility of the top-down dialogue model can only be made upon empirical evaluation, a user study is necessary to validate the proposed model.

In addition to analysis of free-form user feedback, evaluation of a dialogue model can be automated by inspecting dialogue patterns in the collected dialogue transcripts. In these settings, dialogues can be treated as iterative processes whose key patterns allow us to discern strengths and weaknesses of the dialogue model. To analyse dialogues as processes, we propose a use of process mining techniques.

Process mining is the subfield of data science that aims to provide tools for discovering insights into operational processes and thus supports process improvements [76]. Following the process mining terminology [50], an instance of a process (i.e., a specific explanatory dialogue) is denoted as a *trace* τ.

Table 4
An example of an event log (the activities in bold are those produced by the system; the user-produced activities are those in italics)

Case	*Activity*	*Start*	*End*
Dialogue_1	**claim**	2022-06-09 11:54:12	2022-06-09 11:54:12
Dialogue_1	*why-explain*	2022-06-09 11:54:12	2022-06-09 11:54:21
Dialogue_1	**explain-f**	2022-06-09 11:54:21	2022-06-09 11:54:22
Dialogue_1	*what-details*	2022-06-09 11:54:22	2022-06-09 11:54:42
Dialogue_1	**elaborate**	2022-06-09 11:54:42	2022-06-09 11:54:42
Dialogue_1	*why-not-explain*	2022-06-09 11:54:42	2022-06-09 11:55:58
Dialogue_1	**explain-cf**	2022-06-09 11:55:58	2022-06-09 11:56:00
Dialogue_1	*what-details*	2022-06-09 11:56:00	2022-06-09 11:56:32
Dialogue_1	**elaborate**	2022-06-09 11:56:32	2022-06-09 11:56:33
Dialogue_1	*accept-u*	2022-06-09 11:56:33	2022-06-09 11:57:28
Dialogue_1	**accept-s**	2022-06-09 11:57:28	2022-06-09 11:57:28
Dialogue_2	**claim**	2022-06-15 17:03:34	2022-06-15 17:03:34
Dialogue_2	*why-explain*	2022-06-15 17:03:34	2022-06-15 17:04:22
Dialogue_2	**explain-f**	2022-06-15 17:04:22	2022-06-15 17:04:23
Dialogue_2	*what-is*	2022-06-15 17:04:23	2022-06-15 17:04:50
Dialogue_2	**clarify**	2022-06-15 17:04:50	2022-06-15 17:04:50
Dialogue_2	*why-not-explain*	2022-06-15 17:04:50	2022-06-15 17:05:38
Dialogue_2	**explain-cf**	2022-06-15 17:05:38	2022-06-15 17:05:40
Dialogue_2	*why-not-alternative*	2022-06-15 17:05:40	2022-06-15 17:06:12
Dialogue_2	**alter-cf**	2022-06-15 17:06:12	2022-06-15 17:06:13
Dialogue_2	*what-details*	2022-06-15 17:06:13	2022-06-15 17:06:59
Dialogue_2	**elaborate**	2022-06-15 17:06:59	2022-06-15 17:07:00
Dialogue_2	*reject-u*	2022-06-15 17:07:00	2022-06-15 17:07:49
Dialogue_2	**reject-s**	2022-06-15 17:07:49	2022-06-15 17:07:49

Subsequently, each trace consists of the set of *activities* A (in this case, locutions). In turn, a specific instance (realisation) of an activity $\alpha \in A$ (i.e., a dialogue move) is referred to as an *event* ε. Altogether, a collection of explanatory dialogues makes up the so-called *event log*.

An example of an event log basing on a collection of explanatory dialogues is depicted in Table 4. It contains two traces (i.e., Dialogue_1 and Dialogue_2) that represent instances of the recorded explanatory dialogues between (possibly, different) user(-s) and the given system (i.e., an interpretable rule-based classifier). In total, the process model contains 22 events each of which is essentially a specific dialogue move paired with the corresponding locution. Figure 6 illustrates the corresponding process model graph. The visual representation of the process model facilitates detection of the activity patterns (i.e., subprocesses characterising common parts of distinct dialogues) taking place in the collection of dialogues.

A dialogue protocol can be represented as a finite state machine whose nodes are the locutions modelled, edges being legitimate transitions between different states of the dialogue (e.g., from a request to all possible responses). In terms of process mining, one can represent the dialogue protocol as the so-called *process model* – a directed graph $M = \langle N, E \rangle$ where the set of nodes $N \subseteq A \cup \{\text{Start}, \text{End}\}$ is composed of the process activities and the set of edges $E \subseteq N \times N$ represents (possibly, causal) relations between pairs of activities where Start and End are, respectively, the start and end time of execution of the corresponding activity.

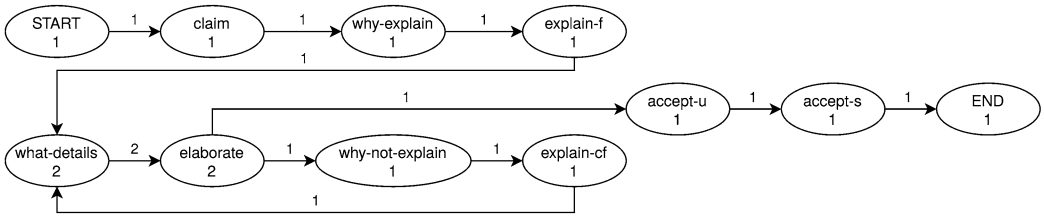


Fig. 6. The graphical view of the process model corresponding to the example Dialogue$_1$ in Table 4.

To analyse the actually recorded dialogues quantitatively, we suggest that the so-called *conformance checking* procedure be applied. In process mining, conformance checking is applied to relate the events in the actually registered processes and the process model in order to identify commonalities and discrepancies between the former and the latter. In the case of evaluating the proposed dialogue game, all the moves made by both dialogue game players follow the previously defined dialogue protocol. Hence, no deviation from the protocol can be observed. Instead, conformance checking allows us to highlight the most (and the least) frequent dialogue patterns in the event log and evaluate it against the process model (i.e., the dialogue protocol). Conformance checking can lead to obtaining data-driven knowledge of the least frequently submitted requests and/or dialogue state transitions, which can be used to modify the originally proposed dialogue protocol in order to increase its quality.

To sum it up, the proposed dialogue model can be evaluated in two complementary ways: qualitatively and quantitatively. On the one hand, qualitative free-form user feedback (e.g., in the form of a post-experiment survey) can point to missing requests or transitions between existing requests in the dialogue protocol. On the other hand, the least frequent dialogue patterns may signal their futility for explanatory purposes of the dialogue model. In process mining, a frequency threshold value can, for example, be set to subsequently optimise the process model by removing the least observed model patterns. Similarly, the least frequent requests or responses may be removed from the dialogue protocol if the empirically grounded threshold value is available and set prior to evaluation. As a result, process mining is shown to serve as a methodological basis for quantitative evaluation of the proposed dialogue model. In combination with free-form user feedback for qualitative evaluation of the dialogue protocol, process mining is able to provide us with further insights w.r.t. the quality of a dialogue model.

5. Experimental settings

In order to evaluate the proposed model of explanatory dialogue following the aforementioned evaluation framework, we carried out an exploratory user study. In the remainder of this section, we describe the setup of the human evaluation study. Thus, Section 5.1 describes the datasets used as the basis for training the classifiers for the study. Section 5.2 outlines technicalities of the explanation generation method used in the given experiment. Section 5.3 outlines the distinctive characteristics of the classifiers trained on the aforementioned datasets. Section 5.4 discusses the stimuli selection as well as the design of the dialogue system used in the experiment.

5.1. Datasets

In our study, we used the following three datasets: basketball player position [3], beer style [13], and thyroid disease diagnosis [19]. All three datasets serve to solve a multiclass classification problem in three different application domains. First, the basketball players position dataset presupposes

five classes related to the following player positions: $Y_{\text{basketball}} = \{$*point-guard, shooting-guard, small-forward, power-forward, center*$\}$. Second, the beer style dataset (as was used in the illustrative example in Section 3.2) categorises instances of beer to belong to one of the following eight classes: $Y_{\text{beer}} = \{$*Blanche, Lager, Pilsner, IPA, Barleywine, Stout, Porter, Belgian strong ale*$\}$. Third, the thyroid disease dataset presupposes the following four potential labels: $Y_{\text{thyroid}} = \{$*no hypothyroid, primary hypothyroid, compensated hypothyroid, secondary hypothyroid*$\}$.

To guarantee consistent and comparable results, only numerical continuous features were used for training the corresponding classifiers. Further, all the features were mapped to linguistic terms as follows. The beer style dataset was annotated by an expert brewer, therefore it contains original feature-value partitions. The features from the other datasets were split in three uniform intervals of equal length, each of which was mapped to the following linguistic terms: $\langle$*low, medium, high*$\rangle$ (except for the feature *height*, which is described with 5 linguistic terms, in the basketball player position dataset). Table 5 summarises information on the features from all the datasets as well as the corresponding linguistic terms, with the numerical intervals attached.

5.2. Explanation generation method

To evaluate the dialogue game proposed in this paper as a communication interface between the system and the user, we generate multiple factual and CF explanations using the *XOR* method [72]. This explanation generation method operates on the rule base (i.e., a set of decision paths to each class) of a rule-based interpretable classifier (e.g., a fuzzy rule-based classification system or a decision tree DT where branches are first transformed into a list of rules). All automatic explanations follow the structure of the decision path (in the case of the factual explanation) or the minimally different decision path leading to the given CF class (in the case of the CF explanation). The following pipeline of four steps constitutes the explanation generation process:

(1) **Rule vectorisation.** Each rule found in the rule base is represented as a (binary, in the case of the XOR method) vector of all possible feature-value pairs. In the case of a DT, the values of the vector are all the unique conditions (e.g., "bitterness $\leqslant$ 10") found in the set of DT nodes.
(2) **Relevance estimation.** Once the rules are vectorised, a distance is calculated between vectors representing the decision path vector (responsible for the prediction) and each rule leading to the given (factual or CF) class. In the case of the XOR method, the exclusive-OR function calculates the distance between the vectors. The vectors are then ranked in accordance with the distances. The minimally distant rule is selected as a template for the output explanation following the conventional definition of a CF explanation.
(3) **Linguistic approximation.** Each interval found in the selected rule is mapped to the predefined linguistic terms by measuring the similarity between the set of numerical values corresponding to this interval and each set of numerical values for the corresponding feature. The most similar linguistic term is selected for the given feature.
(4) **Surface realisation.** The linguistically approximated rule is passed on to the surface realisation module that outputs a template-based grammatically correct high-level explanation. Similarly, the corresponding numerical intervals are used to generate a low-level explanation.

For DTs, factual explanations are essentially the feature-value intervals aggregated along the decision path. This explanation generation method presupposes that alternative factual explanations cannot be generated because alternative decision paths leading to the same predicted class would not adequately

Table 5
Numerical intervals of the features as well as the corresponding linguistic terms

Feature	*Linguistic term*	*Range of values*
Height	Short	[1.810, 1.888]
	Medium-height	[1.888, 1.966]
	Tall	[1.966, 2.044]
	Very tall	[2.044, 2.122]
	Extremely Tall	[2.122, 2.200]
Minutes	Low	[8.410, 14.290]
	Medium	[14.290, 20.160]
	High	[20.160, 26.040]
Points	Low	[2.800, 6.200]
	Medium	[6.200, 9.600]
	High	[9.600, 13.000]
2-points field points percentage	Low	[34.400, 45.500]
	Medium	[45.500, 56.600]
	High	[56.600, 67.700]
3-points field points percentage	Low	[0.000, 15.170]
	Medium	[15.170, 30.330]
	High	[30.330, 45.500]
Free throws	Low	[43.900, 59.300]
	Medium	[59.300, 74.700]
	High	[74.700, 90.100]
Rebounds	Low	[1.600, 3.330]
	Medium	[3.330, 5.070]
	High	[5.070, 6.800]
Assists	Low	[0.200, 1.930]
	Medium	[1.930, 3.670]
	High	[3.670, 5.400]
Blocks	Low	[0.000, 0.570]
	Medium	[0.570, 1.130]
	High	[1.130, 1.700]
Turnovers	Low	[0.200, 0.630]
	Medium	[0.630, 1.070]
	High	[1.070, 1.500]
Global assessment	Low	[4.000, 8.370]
	Medium	[8.370, 12.730]
	High	[12.730, 17.100]

(a) Basketball player position

Feature	*Linguistic term*	*Range of values*
Colour	Pale	[0.000, 3.000]
	Straw	[3.000, 7.500]
	Amber	[7.500, 19.000]
	Brown	[19.000, 29.000]
	Black	[29.000, 45.000]
Bitterness	Low	[7.000, 21.000]
	Low-medium	[21.000, 32.500]
	Medium-high	[32.500, 47.500]
	High	[47.500, 250.000]
Strength	Session	[0.035, 0.052]
	Standard	[0.052, 0.067]
	High	[0.067, 0.090]
	Very high	[0.090, 0.136]

(b) Beer style

Feature	*Linguistic term*	*Range of values*
Thyroid-stimulating hormone (TSH)	Low	[0.000, 3.333]
	Medium	[3.333, 6.666]
	High	[6.666, 10]
Triiodothyronine (T3)	Low	[0.050, 3.560]
	Medium	[3.560, 7.080]
	High	[7.080, 10.060]
Total thyroxine (TT4)	Low	[2.000, 94.660]
	Medium	[94.660, 187.330]
	High	[187.330, 280.000]
Thyroxine utilization (T4U)	Low	[0.250, 7.900]
	Medium	[7.900, 15.550]
	High	[15.550, 23.200]
Free thyroxine (FTI)	Low	[2.000, 84.660]
	Medium	[84.660, 167.330]
	High	[167.330, 250.000]

(c) Thyroid disease

explain the exact reasoning of the DT for the given test instance. On the contrary, alternative CF explanations are considered for explaining hypothethical, non-predicted outcomes. Once the explainer generates an explanation, it is then passed on to dialogue system upon request.

Table 6
Main characteristics of the datasets and the corresponding classifiers used in the experiments

Dataset	*# of instances*	*Accuracy*	*Precision*	*Recall*	*F-score*
Basketball	50	54.000%	0.535	0.540	0.529
Beer	400	93.500%	0.936	0.935	0.935
Thyroid	3772	95.334%	0.947	0.953	0.948

Table 7
Number of decision paths and CF classes for each dataset under consideration

Dataset	*Class*	*# of decision paths*	*# of alternative CF explanations*
Basketball	Point-guard	2	1
	Shooting-guard	2	1
	Small-forward	3	2
	Power-forward	3	2
	Center	4	3
Beer	Blanche	1	–
	Lager	2	1
	Pilsner	6	5
	IPA	8	7
	Barleywine	4	3
	Stout	2	1
	Porter	4	3
	Belgian strong ale	1	–
Thyroid	No hypothyroid	220	219
	Primary hypothyroid	49	48
	Secondary hypothyroid	2	1
	Compensated hypothyroid	186	185

5.3. Classifiers

In our human evaluation study, we use DTs as classifiers. Notably, DTs offer interpretable rule-based explanations that can be retrieved from their readily available internal structure. Three variants of DTs (*J48, RandomTree, REPTree*) were generated using the data mining tool Weka [30] and inspected for all the considered datasets. All the DTs were trained using 10-fold cross-validation.

It turns out that only the *RandomTree* algorithm generates at least two decision paths to all the classes in all the datasets under consideration (except for classes *Blanche* and *Belgian Strong Ale* in the beer style dataset). First, this guarantees the existence of at least one CF explanation for any class in each dataset for any test instance selected. Subsequently, it provides at least one alternative explanation for the given CF class. Since the other inspected DT algorithms did not provide at least one alternative CF explanation for the considered datasets, the *RandomTree*-based DTs were selected for all the use cases as classifiers whose predictions were to be explained in the study. Table 6 summarises main characteristics of the DTs used in the human evaluation study. Table 7 indicates numbers of decision paths for each CF class for each dataset.

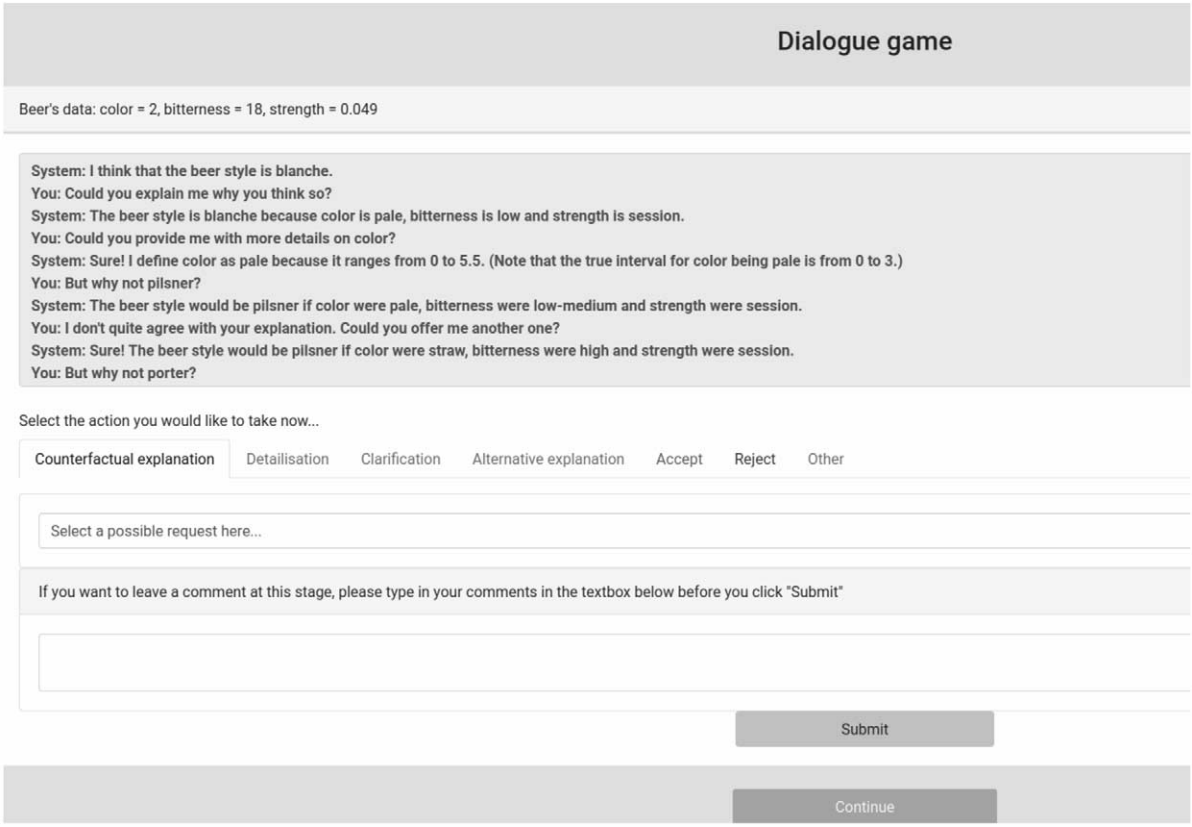


Fig. 7. An example of a dialogue game human evaluation survey (the beer style dataset scenario).

5.4. *Online evaluation settings*

In order to execute human-machine interaction governed by means of the dialogue game proposed, we designed and implemented an online evaluation system. The corresponding ethical considerations are outlined in Appendix A. Figure 7 presents an example screen of the implemented software tool.[3] Further, the source code of the dialogue game survey, the DTs used in the experiments, and the collected experimental data are made publicly available.[4]

In the course of the study, the participants were presented the characteristics of a test instance following the chosen scenario (dataset). The participants did not have any prior knowledge about the dataset. They were asked to interact with the system until they could make an *informed* decision on acceptance or rejection of the system's claim. The participants determined the flow of the dialogue, as they requested necessary information to make a final decision.

Three test instances (one per dataset) were selected so that they would represent correctly predicted real data. Table 8 outlines the characteristics of the test instances used in the study. The following factual explanations were generated for the considered test instances:

[3] https://tec.citius.usc.es/dialgame

[4] https://gitlab.citius.usc.es/ilia.stepin/fcfexpgen (branch "dialgame").

Table 8
Test instance characteristics

Height	*Minutes*	*Points*	*2-points field goals percentage*	*3-points field goals percentage*	*Free throws*	*Rebounds*	*Assists*	*Blocks*	*Turnovers*	*Global assessment*	*Class*
1.85	21.19	9.2	43.1	40.0	81.9	1.9	3.8	0.0	0.7	8.8	Point-guard

(a) Basketball player position

Colour	*Bitterness*	*Strength*	*Class*
2	18	0.049	Blanche

(b) Beer style

Thyroid-stimulating hormone (TSH)	*Triiodothyronine (T3)*	*Total thyroxine (TT4)*	*Thyroxine utilization rate (T4U)*	*Free thyroxine index (FTI)*	*Class*
4.6	1.2	48	0.89	54	Secondary hypothyroid

(c) Thyroid diagnosis

- **Basketball:** "The player's position is point-guard because the number of rebounds is low and the number of assists is high."
- **Beer:** "The beer style is Blanche because colour is pale, bitterness is low and strength is session."
- **Thyroid:** "The patient has secondary hypothyroid because thyroid-stimulating hormone is medium, triiodothyronine is medium and total thyroxine is low."

Similarly, all the high-level automatically generated CF explanations contained only textual descriptions of the features involved. As all the features are numerical (either integer or real-valued), responses to detailisation requests would provide subjects with intervals to which the linguistic terms are mapped. Further, the users were then informed about the classifier's numerical intervals found for the given feature along the given decision path. These details were assumed to facilitate matching the system's claim with the feature-value pairs of the test instance.

Noteworthy, the same study participants could select multiple datasets to play the dialogue game. Therefore, the numbers of records for each dataset do not represent unique users. For this reason, whenever we hereinafter mention the study participants (subjects), we refer to the actually collected transcripts of explanatory dialogues.

Upon completion of the experiment, the study participants were asked to optionally provide their demographic data and leave free-text responses to the following questions and/or suggestions:

Q1 "If you could add other types of requests to the system, what would those be?";
Q2 "Did the interaction with the system change your initial (dis-)belief in the system's prediction? Why (not)?";
Q3 "If you have any other comments for us, please leave them in the textbox below."

Last but not least, all the collected dialogue transcripts were transformed into event logs. On the basis of the event logs, process models were then constructed for each use case. In addition, a global process model of all the event logs was calculated.

Table 9
General properties of the collected dialogues

Property	Dataset			All datasets
	Basketball	Beer	Thyroid	
		Number of dialogue moves		
Mean	12.57	15.76	10.11	14.17
Median	12.00	15.00	9.00	13.00
St.dev.	6.98	7.00	3.18	6.83
		Time taken (min)		
Mean	04 m 09 s	08 m 47 s	05 m 17 s	07 m 10 s
Median	04 m 01 s	05 m 42 s	04 m 54 s	04 m 35 s
St.dev.	01 m 39 s	09 m 39 s	02 m 31 s	07 m 55 s

6. Experimental results

In this section, we report the collected human evaluation results. Section 6.1 presents the quantitative results of the study (i.e., descriptive analytics of the collected dialogues and insights from the process models). Section 6.2 reports the qualitative results of the evaluation study (i.e., the free-form feedback that the study participants left optionally after their interaction with the dialogue system).

6.1. Dialogue analytics

A total of 60 dialogue transcripts have been collected in the course of the empirical study. In particular, 14 (23.33%) of the records relate to the basketball player position dataset. In turn, 37 (61.67%) transcripts are composed as the result of interaction with the classifier trained on the beer style dataset. In addition, 9 (15.00%) records reflect user interaction with the thyroid dataset-based classifier. All the collected dialogue transcripts were converted into event logs. The event logs were subsequently used to generate two process models: (1) the one related to the main building blocks of the modelled explanatory dialogue (i.e., claim, explanation, and termination) and (2) the one covering all the locutions produced by the study participants. Process model (1) gives a high-level overview of the user behaviour whereas process model (2) provides insights w.r.t. specific moves made by the study participants.

On average, it took the dialogue game participants around 14 moves for the users to make their final decision with respect to the system's claim. As for the time taken to complete the dialogue game, the study participants spent about 7 minutes to either accept or reject the claim. Table 9 reports average numbers of dialogue moves and the time taken to complete the dialogue for each dataset under consideration.

Figure 8 illustrates the process model corresponding to the three main building blocks of the proposed dialogue game (i.e., claim, explanation, and termination). Thus, all but three participants required (at least, factual) explanation for the given prediction. Almost all of them eventually accepted the system's claim. In the remainder of this section, we are analysing only those transcripts where explanations were requested.

Figure 9 depicts the process model of the collection of explanatory dialogues that displays all the locutions produced. Thus, 331 explanation-related requests (all those covered by the EXPLANATION non-terminal in EDG) have been registered from the 57 participants who required explanation for the system's claim. The edge labels for the explanation-related requests in Fig. 9 show that the study participants actively exploited all the explanation-related requests that were designed in the original protocol.

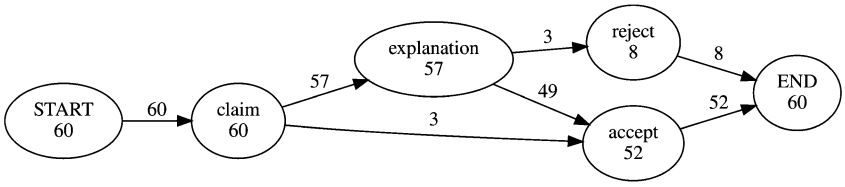


Fig. 8. The process model of all the collected explanatory dialogues based on the main EDG building blocks. The block "termination" is split into "accept" and "reject".

On the one hand, a majority of the participants submitted further explanation-related requests (in this case, detailisation or clarification) upon receiving the factual explanation. On the other hand, a quarter of all the study participants considered the factual explanation sufficiently comprehensive to immediately request a (set of) CF explanation(-s).

The locution-level process model (see Fig. 9 for details) allows us to observe the answers to which requests were the most decisive for the participants to make their final decisions. Thus, the system's claim was mainly accepted immediately after CF explanations (including those alternative) were presented whereas only one participant accepted the system's claim did so as soon as the factual explanation was offered. The other explanation-related requests (i.e., detailisation and clarification) are found to have contributed less to immediate acceptance of the system's claim. As for claim rejections, alternative CF explanations happen to most frequently trigger negative user decisions. Notably, alternative CF explanations were requested for nearly a half of all 76 CF explanations offered. In most cases, study participants stopped exploring the explanation space for the given CF class after the second-best ranked CF explanation was offered. However, third-best ranked CFs were requested to a limited extent.

It is worth noting that further insights into the quantitative results for individual use cases can be found in Appendix D.

6.2. *User feedback*

In this section we present all the free-form comments that the study participants left upon finishing their interaction with the system and summarise the most informative of them. Recall that study participants were encouraged to leave answers to two questions (Q1 and Q2) and/or indicate their free-form suggestions (Q3) unrelated to Q1 or Q2 after their interaction with the implemented dialogue system. The collected responses to Q1–Q3 are presented in Tables 10–12. As all the comments shown are original, some may contain grammatical, lexical, and/or orthographic errors. All the users' statements are codified as follows: "$Cx.y$" where C stands for "comment", x is the corresponding question number and y is the answer number.

Table 10 presents all the answers to Q1 ("If you could add other types of requests to the system, what would those be?") that we collected throughout the study. Two comments (C1.1 and C1.2) are related to the basketball player position. Six statements (C1.3–C1.8) were made as a result of interaction with the system in the beer style case settings. One study participant left his or her comment (C1.9) after playing with the thyroid disease diagnosis scenario.

Regarding Q1, the study participants would like to extend the actual dialogue model so that it could inform them about the second most probable decision, or the technicalities of the decision-making system (e.g., the accuracy of the system). In addition, further definitions of notions related to the domain knowledge (see Comment C1.6, Table 10) were desired. Notably, concerns were raised about the in-

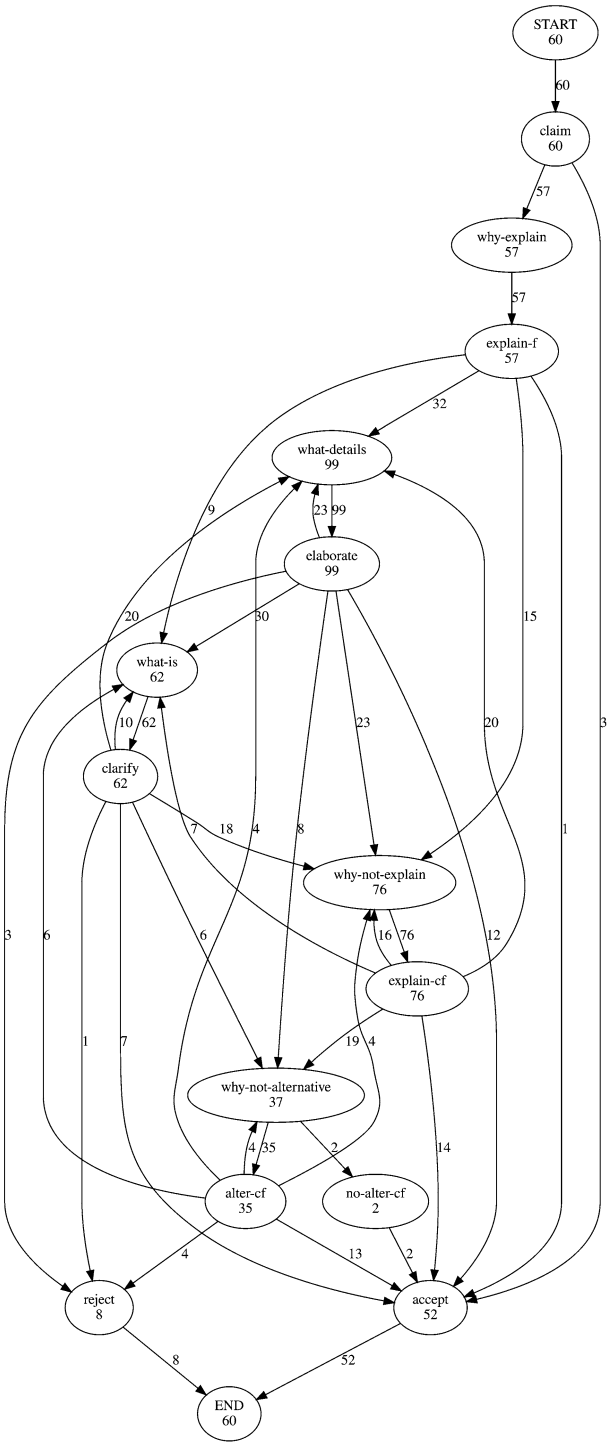


Fig. 9. The full process model of all the collected explanatory dialogues. For illustrative purposes, pairs of termination nodes, i.e. *{accept-u, accept-s}* and *{reject-u and reject-s}*, are merged into *accept* and *reject*, respectively.

Table 10
Study participants' answers to Q1 ("If you could add other types of requests to the system, what would those be?")

ID	User's statement
C1.1	"I'm unsure"
C1.2	"explain what is your primary goal for the predictions you are making"
C1.3	"Summarisation"
C1.4	"In clarifications, I'd like to not only get the definition of the strength but also the types of strength that exist. For example, Blanche's strength is session but I have no idea what session means."
C1.5	"It would be good to have some clarification of different terms than fixed one like color"
C1.6	"I would add more elaborated set of definitions, i.e. definitions of technical terms which are used for definitions."
C1.7	"how did you measure the (.); what is the accuracy of this measurement tool? What is the probability of your prediction?; how did you calculate this probability?"
C1.8	"I would like the possibility of going back to previous points. It seems to me that after the counterfactual explanation I was stuck on it, and going back to the original prediction was at least not intuitive. A graph of the history of dialogue that would allow me to travel through explanations would be great. Predefined options were not very clear to me I think a better explanation with examples would be beneficial. There might be corner cases on different topics that would make differentiating those options even harder."
C1.9	"Second most probable choice (differential diagnoses in the case of the thyroid case)"

ability to post-process the pieces of explanation that had already been discussed (see Comment C1.8, Table 10).

Table 11 shows all the collected answers to Q2 ("Did the interaction with the system change your initial (dis-)belief in the system's prediction? Why (not)?"). Five study participants (C2.1–C2.5) answered Q2 after making their decision on the automatic basketball player position classification. Ten statements (C2.6–C2.15) were made as a result of interaction with the system in the beer style case settings. Two study participants (C2.16–C2.17) commented on their interaction with the system, as the thyroid disease classification scenario was executed.

Regarding Q2, a fair number of commentators found the offered automated explanations convincing and satisfactory. Comment C2.5 (Table 11) illustrates that this was, in part, achieved due to the possibility to opt for factual explanations. In addition, some study participants positively assessed the ability to query the system for CF explanations (see Comment C2.8, Table 11) and further details and clarifications (see Comment C2.3, Table 11). Some of the commentators whose initial (dis-)belief in the system's claim did not change in the course of their interaction with the system remarked that the explanations offered were nevertheless satisfying (see Comment C2.2, Table 11) and supportive enough w.r.t. the system's claim (see Comment C2.11, Table 11).

Table 12 presents all users' free-form suggestions (Q3: "If you have any other comments for us, please leave them in the textbox below."). One comment (C3.1) was left after a dialogue with system w.r.t. the basketball player position classification whereas two statements (C3.2–C3.3) were made as a result of interaction with the system in the beer style case settings.

Regarding Q3, one study participant commented that the system's responses were too fast (see Comment C3.1, Table 12). In addition, another participant pointed out the need for supportive visualisation tools, a clearer distinction between detailisation and clarification requests, and different structures for alternative explanations for the same CFs (see Comment C3.2, Table 12). Finally, predictions for other data instances are found desired to be inspected to develop big picture thinking about the reasoning of the system (see Comment C3.3, Table 12).

Table 11

Study participants' answers to Q2 ("Did the interaction with the system change your initial (dis-)belief in the system's prediction? Why (not)?")

ID	User's statement
C2.1	"Yes. It provided a counter argument of why they had provided that prediction specifically and not another that I suggested."
C2.2	"No because the system had the numbers, so I believed it from start to finish."
C2.3	"I have no knowledge of basketball but the explanations were convincing so I was happy to accept the prediction after asking further questions"
C2.4	"It made me feel that the system has a certain etos but did not teach me about how these predictions are actually computed"
C2.5	"The system was able to successfully convince me of the prediction based on the factual information it provided."
C2.6	"No"
C2.7	"It didn't describe the details of the low bitterness when I asked about bitterness following a discussion about ipa. It provided me with details about high bitterness and outlined that ipa has high bitterness. I could not clarify the bitterness low level range that was the suggested prediction of Blanche."
C2.8	"Yes, seeing the classifications of the other types that is suspected made me accept that this prediction must be correct"
C2.9	"Yes, it gave me a deeper understanding of beer classification. It is a nice way to learn and to gain trust in AI system."
C2.10	"The system responses were good and straight to the point so it was quite convincing."
C2.11	"It did not. I thought it was pretty accurate from the start and given the example before the experimental item I could already gather a good idea of what was expected."
C2.12	"yes, in the beginning I didn't understand one of the words and my first thought was that the word, which was awkward to me, was an effect of system's malfunctioning."
C2.13	"I did not have a strong initial belief about the system prediction. However, it was convincing enough for me."
C2.14	"No – I had no experience or grounds on which to doubt what I was being told. The questions and answers seemed a matter of technical specification and not a matter of beliefs."
C2.15	"Not really, I know it is difficult for an AI system to have long dialogues as it needs to take account with everything that has been said before."
C2.16	"Not really, because I didn't have any expectations"
C2.17	"Clarification of the prediction terms as well as the features would be useful. For example, what hypothyroid means etc"

7. Discussion

The findings reported in the previous section enable us to outline several remarkable observations. As expected, high numbers of detailisation and clarification requests have been registered from the users interacting with a classifier in the settings where they did not have any prior knowledge of the dataset that the classifier had been trained on. As the users started their interaction with the system only having feature-value pairs of the test instance at their disposal, they oftentimes required not only an explanation to the system's claim but, perhaps, more importantly, definitions of the features that made part of the explanation or the numerical ranges over which the features were defined. The fact that a high number of requests for alternative explanations have been registered across all the use cases confirms that the most relevant explanation from the system's point of view may be far from the most relevant (or satisfactory) from the user's point of view.

As the same prediction can be explained in different ways, it turns out to be particularly important to extend the protocol so that it does not only offer the opportunity to rephrase the initially offered explanation but also enables the system to send requests to the user. For instance, if two pieces of

Table 12
Study participants' suggestions w.r.t. to Q3 ("If you have any other comments for us, please leave them in the textbox below")

Comment ID	User's statement
C3.1	"The responses were very fast, a slight delay after receiving a request would improve how the answer appears"
C3.2	"In the beginning, it'd be nice to have some kind of photo prompt together with the beer data to help vizualise what we are talking about. It's a bit hard to distinguish between detalisation and clarification. I didn't see the difference in the structures of counterfactual explanation and alternative explanation. In my case, for the counterfactual explanation, I asked about pilsner and when giving me an alternative explanation the system also used pilsner so I didn't get new information from the last request."
C3.3	"I would be curious to learn more about other topics and other predictions on the subject I took (in this case, beer)."

explanation are deemed equally relevant by the explanation generation module, requiring additional information from the user about his or her preferences may be crucially important for successful fine-tuning of the explanation being processed. On the one hand, both such explanations can be presented simultaneously. Then, the user is to decide the format and/or ordering of the output explanations. On the other hand, the system can submit a request to the user to infer the actual user's needs taking into consideration the known differences between two explanations.

The qualitative results of the human evaluation study allow us to suggest a number of empirically-driven critical questions (CQ) to the system's prediction. Recall that our factual and CF textual explanations (in the simplest form) follow the templates "The test instance is [CLASS] because [FEATURE] is [VALUE]" and "The test instance would be [CLASS] if [FEATURE] were [VALUE]", respectively. We can therefore address CQs both to the prediction (variable CLASS in the example above) and to (components of) the explanation (the variables FEATURE and VALUE in the example above). Driven by the registered user feedback, the prediction-related CQs (CQ1, CQ2, and CQ3) can be exemplified as follows:

CQ1 Is the system's prediction correct?
CQ2 What is/are the accuracy/precision/recall/F-score of the system that predicted [CLASS]? (following C1.7 from Table 10);
CQ3 How were the accuracy/precision/recall/F-score calculated? (following C1.7 from Table 10).

In turn, the features and values of the given explanation may give rise to explanation-related CQs. For example, the feature values may be subject to explanation-related CQs that may occur when processing responses to detailisation requests (CQ4 and CQ5) while the definitions of the features themselves may be questioned upon performing clarification requests (CQ6):

CQ4 What data justify [VALUE] for [FEATURE]? (in the case of high-level explanations);
CQ5 Is [VALUE] consistently defined for [FEATURE] in [INTERVAL]? (where [VALUE] is the linguistic term of some high-level explanation's feature and [INTERVAL] is the corresponding numerical interval of the low-level explanation);
CQ6 Is the source of information of the definition of [FEATURE] credible?

The proposed dialogue model has a number of limitations. As it can be applied directly only to interpretable rule-based classifiers enhanced with explainers providing textual explanations, the communication between the system and the user may appear overly restricted. In light of the assumptions made in Section 2, parts of the protocol may have to be adjusted when dealing with, for example, categorical variables or a poorly interpretable feature space. In addition, the structure of the protocol may have to

be made more flexible, as handling the previously processed explanations (for example, those for other CF classes) is not permitted.

Remarkably, the set of locutions included in the presented protocol is by no means exhaustive. The qualitative results of the human evaluation study signal a number of desired extensions to the proposed dialogue model. The users would, for example, appreciate to know more about the definitions of the linguistic terms. The modular architecture of the EDG production rules allows for adapting the dialogue game for developer's as well as user's needs. In this regard, the clarification requests can be made applicable not only to the features themselves but also to the values of the linguistic variable that appear in high-level explanations as well as domain knowledge-related terms. In addition, the proposed dialogue protocol might as well incorporate visual information (e.g., pictures of the domain knowledge available upon request) for detailisation requests.

8. Related work

A variety of computational argumentation models have proven to be efficient tools for explanatory dialogue modelling in the context of XAI. For instance, Arioua et al. [4] propose a formal model of argumentative explanatory dialogue to acquire new knowledge in inconsistent knowledge bases. Calegari et al. [10] implement a mechanism of reasoning over defeasible preferences using elements of abstract and structured argumentation. Groza et al. [26] model explanatory dialogues combining rule-based arguments extracted from both ML classifiers and expert knowledge in favour or against a given classification of retinal disorder. Subsequently, the arguments are used to persuade the other parties in multi-agent system settings.

Argumentative explanatory dialogues are of particular interest among XAI researchers, as they provide means for customisation of automated CF explanations in light of the collected user feedback [70]. There exist a large number of distinct techniques that allow for integrating user feedback to personalise initially generated CF explanations. For example, Suffian et al. [73] operate on user's preferred features and the corresponding ranges of values to fine-tune the originally generated explanation. Their FCE method first generates synthetically a set of CF data points where the preferred features range in the selected intervals. Then, the model aims to detect the most relevant (yet personalised) CF by searching for the minimally different (in terms of distance) CF data point from the generated synthetic data. Behrens et al. [6] propose a dynamically updated framework for user-specific explanation generation for knowledge graphs. More precisely, the user expresses his or her preferences by selecting two desired sets of graph nodes and, subsequently, ordering the selected generated meta-paths (i.e., sequences of alternated nodes and edges). Ghazimatin et al. [24] collect user feedback on explanations themselves for a recommender system to improve its performance. In this case, the user feedback is essentially a binary value signalling the similarity of an explanation to the recommendation. De Toni et al. [18] consider the problem of causal CF explanation generation as algorithmic recourse (i.e., overturning unfavourable ML-based model's prediction). In their reinforcement learning-based model, the user is asked to choose the best subsequent action from the so-called "choice set". The user's responses are then used to optimise the model's weights via Bayesian estimation and update the user's state.

Early computational models of explanatory dialogue stress that the context of explanation should depend on user's familiarity with the concepts presented to him or her [14]. Further, the end user is argued to necessarily build a sound mental model of the system to successfully interact with it [84]. However,

only a few of argumentative explanatory dialogue implementations allow for direct dialogic interaction between an AI-based system and a given user for explanation customisation. Despite little evidence, human evaluation of the automatically generated explanations may lead to groundbreaking conclusions. For instance, Rago et al. [56] emphasise the need for multi-modality of the generated argumentative explanations, as users are found to generally prefer tabular explanations over textual ones but also textual over conversational. In addition, explanations containing a greater number of features (aspects) are, in general, found to be preferred.

Formal dialogue games provide an intuitive transparent tool of information exchange between the agents involved [54]. They have been extensively used in a wide range of AI applications, such as multi-agent systems [44] and recommendation systems [42]. Dialogue games have shown to have great potential for explanatory dialogue modelling [36]. The first dialogue games for (computational) explanatory dialogue modelling trace back to works by Walton [81] and Modgil and Caminada [46]. Arioua and Croitoru [5] propose a dialogue game to formalise Walton's dialectical system of explanatory dialogues. However, their formalism does not take into account some key properties of explanation (contrastive, selected, and social) as well as user-specific needs addressed in the field of XAI. On the other hand, Shao et al. [67] explain a neural network's classification output enabling the user to adjust the classifier's prediction by enabling the user to prove feedback on the arguments correcting the prediction. Shaheen et al. [65] design two dialogue game-based protocols for generating and communicating explanations for satisfiability modulo theory (SMT) solvers. Thus, their approach distinguishes between a passive explanatory dialogue game where the explainee only inquires explanation and an active game where the user is explicitly asked to confirm or refute the system's assertion. Unfortunately, both protocols lack any empirical evaluation. Alternatively, Sklar and Azhar [69] perform a user study to evaluate a dialogue game-based framework for making cooperative actions in the treasure hunt game. They show that explanations communicated using a dialogue game-based communication protocol lead to above-average user satisfaction. Shams et al. [66] design a dialogue game to explain and justify the best agent's plan in normative practical reasoning settings. Finally, argumentative dialogue game-based models have been proposed for generating model-agnostic local explanations to justify given predictions [55]. To the best of our knowledge, no other dialogue games (including those aforementioned) have ever been evaluated (quantitatively) using process mining techniques like those introduced in this paper.

The previously mentioned protocols were mainly proposed for modelling information-seeking or inquiry explanatory dialogues. However, the formalism of dialogue games is also suitable for (and extensively applied to) modelling persuasive explanatory dialogues. Thus, Sassoon et al. [61] center explanatory dialogue around instances of a domain-specific argumentation scheme guided with the corresponding critical questions. Depending on the degree of agreement between the agents, the explanatory dialogue is then modelled in one of the three following modalities: information-seeking, deliberation, or persuasion. Morveli-Espinoza et al. [48] propose a protocol for persuasive negotiation dialogues where agents exchange explanatory and rhetorical arguments. Similarly to our approach, they consider alternative responses to be, in part, attacks to the previously uttered arguments. However, their protocol does not tackle CF explanations.

Last but not least, a large body of research has attempted to formalise dialogue by means of dialogue grammars [32,58]. Thus, they have been regarded as a natural interface between the underlying speech acts and actually produced utterances [64]. Dialogue grammars have been shown to dis-

ambiguate between distinct dialogue flow patterns (e.g., elaboration, digression, problem resolution, to name a few) [33]. In addition, dialogue grammars facilitate induction of task-based dialogue systems [22]. Beneficially, such grammars can be learned from dialogic data in an unsupervised manner [23]. Further, dialogue grammars are scalable yet universally induced from any domain [38]. Subsequently, the grammar-based approach to dialogue modelling has been enhanced with methods of corpus-based query generation for natural language understanding [34].

Dialogue grammars are found to model human-human dialogue [68] as well as human-machine dialogue [39]. Thus, dialogue grammars appear particularly useful for multimodal human-machine interaction. For instance, hybrid multiset grammars are proposed to govern speech and textual input jointly [20]. On a similar note, Kottur et al. [40] propose a dialogue grammar for visual co-reference resolution. In contrast to the aforementioned approaches where the explanatory dialogue is formalised by means of dialogue grammars, our EDG allows for producing natural language output only. However, a high degree of modularity that dialogue grammars offer makes it possible to extend the dialogue model so that it also outputs visual data (e.g., saliency maps) if such visual explanations are included in the set of terminals of the grammar.

9. Conclusions and future work

In this paper, we presented a new approach for explanatory dialogue modelling. Namely, we designed a dialogue game for the task of communicating explanations for predictions of interpretable rule-based classifiers. Unlike previous approaches, the dialogue protocol proposed in this work allows for effective communication of both factual and CF explanations for expert and lay users. The protocol offers a transparent means of conveying personalised textual rule-based explanations. Its use can be extended to other interpretable rule-based classifiers (e.g., other DT algorithms or fuzzy rule-based classification systems).

Subsequently, we validated the dialogue protocol by carrying out a human evaluation study. The quantitative results (i.e., the reconstructed process models) confirm the necessity in all the proposed requests for explanatory dialogue between the classifier and its user and therefore proves them indispensable for explanatory dialogue modelling. Thus, detailisation and clarification requests are found particularly useful when natural language explanations are presented in the settings where users have no prior knowledge of the dataset. In addition, end users show a high degree of interest in CF explanations in addition to their factual counterparts. Further, they appear to appreciate the possibility to question the initially offered CF explanations across different application domains. Provided that such CF explanations are generated automatically and presented to the user in accordance with their relevance to the test instance (e.g., the distance from the test instance), the proposed protocol allows the explainer to communicate multiple explanations. Hence, it favours diversity of the offered explanations, which is shown to increase their explanatory power. Moreover, the qualitative results show that the proposed dialogue game appears to be an effective tool to convey appealing explanations which were convincing enough for a good number of users. In this sense, the set of the proposed requests and replies turns out to be a potentially effective tool for measuring the effectiveness of (counter-)factual explanation generation frameworks outputting textual explanations in the course of interaction with end users. Finally, the protocol is flexible enough to be adapted in the near future for estimating the trustworthiness, satisfaction, or persuasive capability of automatically generated explanations while preserving the original structure of the given explanatory dialogue modelled. Nevertheless, the proposed protocol may be found somewhat overrestrictive, as it

does not enable end users to submit explanation-related requests for the pieces of explanation whose processing is considered finalised.

The present piece of research opens the door for several lines of future work. Importantly, the proposed dialogue model should be adapted to handle other types of classifiers including those that do not reveal any interpretable information about their internals. In many settings, knowledge of the feature space is unavailable or hard to interpret. Then, the detailisation requests may result being of little utility unless additionally adapted to the functionality of the given classifier. In addition, we intend to enlarge the argumentative potential of the proposed dialogue model by developing further methods of capturing user's preferences. Further work is also necessary to incorporate explanations of other modalities (e.g., visual) for dialogic communication. Whereas the concept of explanation space may be directly applicable to other settings (e.g., a prediction can be explained by means of different pieces of visual information), this may require redefinition of sub-components of the explanation space.

Another important line of future work consists in extending the actual protocol to incorporate explanations of different content and tasks. For instance, it is of peculiar interest to test the applicability of the dialogue protocol in the settings of regression, recommendation, or planning tasks. Finally, we aim to design and carry out further human evaluation experiments on the trade-off between the limitations of the protocol (e.g., underrepresentated locution types) and the persuasive power of explanations that it communicates. Such experiments (e.g., disabling users to perform specific acts) would allow us to estimate the impact of specific requests and further shape the protocol.

Appendix A. Ethical considerations

All the information collected from the human evaluation study participants was in agreement with the European Union's General Data Protection Regulation (GDPR). In addition, this piece of research has been approved by the Ethics Committee of the University of Santiago de Compostela (Spain). Human evaluation was based solely on non-personal or anonymous data. Further, all the participants gave informed consent confirming the following:

- the participant reached the age of majority;
- participation in the study was completely voluntary;
- participation in the study could be terminated at any time;
- participant's anonymous responses would be used for research purposes in accordance with GDPR.

Appendix B. Dialogue protocol

In our model, any explanatory dialogue is modelled in accordance with the protocol outlined below. Thus, the protocol presupposes the following rules:

(1) **Turntaking.** The system initiates the dialogue, i.e. it makes the move m_1 by claiming the prediction from the domain-specific finite set of all possible predictions $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$ corresponding to the dataset classes. Every subsequent even $(m_2, m_4, \ldots)$ and odd $(m_3, m_5, \ldots)$ moves are made by the user and the system, respectively. Each participant is allowed to produce only one locution at a time.

(2) **User's U allowed moves.**

(a) *why-explain*($\hat{y}$, E): U requests to factually explain $\hat{y}$. The explanation store E remains empty. The system is allowed to respond in either of the following ways:

- *explain-f*($\hat{y}$, E, e_f) iff S is able to produce a factual explanation;
- *no-explain-f*($\hat{y}$, E) otherwise.

(b) *why-not-explain*($\hat{y}$, E, y'): U requests to counterfactually explain why $\hat{y}$ and not y'. E must contain a factual explanation for $\hat{y}$: $E = \{e_f(\hat{y})\}$. The system is allowed to produce either of the following locutions:

- *explain-cf*($\hat{y}$, E, y', e_{cf}) if S is able to produce a CF explanation;
- *no-explain-cf*($\hat{y}$, E, y') otherwise.

(c) *what-details*($\hat{y}$, E[,y'], e, Γ) where $e = e^h_f | e^h_{cf}$: U requests details on a feature Γ used in a previously uttered (factual or CF) high-level explanation e ($\Gamma \in e$). In response, S generates one of the locutions below:

- *elaborate*($\hat{y}$, E[,y'], e, Γ, θ) if S is capable of providing U with details on feature Γ;
- *no-elaborate*($\hat{y}$, E[,y'], e, Γ) otherwise. Note that the parameter y' is optional and passed on iff $e = e^h_{cf}$.

(d) *what-is*($\hat{y}$, E[,y'], e, Ψ) where $e = e^h_f | e^h_{cf} | e^l_f | e^l_{cf}$: U requests a definition of a specific feature Ψ being part of (factual or CF, high- or low-level) explanation e ($\Psi \in e$). The system is allowed to respond using one of the following locutions:

- *clarify*($\hat{y}$, E[,y'], e, Ψ, υ) if S can provide U with such a definition;
- *no-clarify*($\hat{y}$, E[,y'], e, Ψ) otherwise. Note that the parameter y' is optional and passed on iff $e = e^h_{cf} | e^l_{cf}$.

(e) *why-alternative*($\hat{y}$, E, e_f): U disagrees (or is not satisfied) with the offered factual explanation e_f and requires an alternative factual explanation. The system responds producing one of the following locutions:

- *alter-f*($\hat{y}$, E, e_f, e'_f) if S is capable of providing U with a different factual explanation;
- *no-alter-f*($\hat{y}$, E, e_f) otherwise.

(f) *why-not-alternative*($\hat{y}$, E, y', e_{cf}): U disagrees with the offered CF explanation e_{cf} and requires that S provide an alternative CF explanation. The system replies using one of the following locutions:

- *alter-cf*($\hat{y}$, E, y', e_{cf}, e'_{cf}) provided that an alternative CF explanation e'_{cf} can be offered;
- *no-alter-cf*($\hat{y}$, E, y', e_{cf}) otherwise.

(g) *accept-u*($\hat{y}$, E): U accepts the prediction $\hat{y}$. In response, the system generates the fairwell locution *accept-s*($\hat{y}$, E).

(h) *reject-u*($\hat{y}$, E): U rejects the prediction $\hat{y}$. In response, the system generates the fairwell locution *reject-s*($\hat{y}$, E).

(3) **System's S allowed moves.**

(a) *claim*($\hat{y}$, E): S claims prediction $\hat{y}$. The knowledge store K and the explanation store E are initialised to be empty. U is allowed to:

- require a factual explanation (locution *why-explain*($\hat{y}$, E));
- accept prediction $\hat{y}$ without any subsequent explanation (locution *accept-u*($\hat{y}$, E));
- reject prediction $\hat{y}$ without any subsequent explanation (locution *reject-u*($\hat{y}$, E)).

(b) *explain-f*($\hat{y}$, E, e_f): S factually explains $\hat{y}$ with e_f and provides U with its high-level component (recall that $e_f(\hat{y}) = \langle e_f^h(\hat{y}), e_f^l(\hat{y})\rangle$). The factual explanation is added to the knowledge store $K = K \cup e_f(\hat{y})$ and the explanation store $E = E \cup e_f(\hat{y})$. The detailisation and clarification stores are populated with the features making part of the explanation e_f. User U is then allowed to:

- require a CF explanation for some CF class $y' \in CFS$ (locution *why-not-explain*($\hat{y}$, E, y'));
- ask for details on a feature $\Gamma \in DET$ of the factual explanation (locution *what-details*($\hat{y}$, E, e_f, Γ));
- demand a definition of some feature $\Psi \in CLAR$ making part of the factual explanation e_f (locution *what-is*($\hat{y}$, E, e_f, Ψ));
- disagree with the factual explanation e_f for prediction $\hat{y}$ and require an alternative factual explanation (locution *why-alternative*($\hat{y}$, E, e_f));
- accept prediction $\hat{y}$ (locution *accept-u*($\hat{y}$, E));
- reject prediction $\hat{y}$ (locution *reject-u*($\hat{y}$, E)).

(c) *no-explain-f*($\hat{y}$, E): S is unable to factually explain $\hat{y}$. U may nevertheless:

- require a CF explanation for some CF class y' (locution *why-not-explain*($\hat{y}$, E, y'));
- accept prediction $\hat{y}$ (locution *accept*($\hat{y}$, E));
- reject prediction $\hat{y}$ (locution *reject*($\hat{y}$, E)).

(d) *explain-cf*($\hat{y}$, E, y', e_{cf}): S counterfactually explains why $\hat{y}$ and not y' with e_{cf} and provides U with its high-level component (recall that $e_{cf}(\hat{y}, y') = \langle e_{cf}^h(\hat{y}, y'), e_{cf}^l(\hat{y}, y')\rangle$). The CF explanation is added to the knowledge store $K = K \cup e_{cf}(\hat{y}, y')$ and the explanation store: $E = E \cup e_{cf}(\hat{y}, y')$. The CF class y' is then eliminated from the CF class store: $CFS = CFS \setminus \{y'\}$. In response, U is allowed to:

- require details on some feature $\Gamma \in DET$ for the given CF explanation (locution *what-details*($\hat{y}$, E, y', e_{cf}, Γ));
- request a definition of a feature making part of the CF explanation e_{cf} (locution *what-is*($\hat{y}$, E, y', e_{cf}, Ψ));
- disagree with the offered CF explanation and ask for an alternative one for the same CF class (locution *why-not-alternative*($\hat{y}$, E, y', e_{cf}));
- require a CF explanation for another CF class from the CF class store $y'' \in CFS$ (locution *why-not-explain*($\hat{y}$, E, y''));
- accept prediction $\hat{y}$ (locution *accept*($\hat{y}$, E));
- reject prediction $\hat{y}$ (locution *reject*($\hat{y}$, E)).

(e) *no-explain-cf*($\hat{y}$, E, y'): S is unable to counterfactually explain why $\hat{y}$ and not y'. The CF class y' is eliminated from the CF class store: $CFS = CFS \setminus \{y'\}$. User U is allowed to:

- require a CF explanation for another CF class $y'' \in CFS$ (locution *why-not-explain*($\hat{y}$, E, y''));

- disagree with the factual explanation and require an alternative to it provided that it is the only explanation that the explanation store E contains (locution *why-alternative*($\hat{y}$, E, e_{cf})) iff $\nexists\, e_{cf} \in E$;
- accept prediction $\hat{y}$ (locution *accept*($\hat{y}$, E));
- reject prediction $\hat{y}$ (locution *reject*($\hat{y}$, E)).

(f) *elaborate* ($\hat{y}$, E [,y'], e, Γ, θ) where $e = e^h_f | e^h_{cf}$: S provides required details θ on feature Γ of a high-level (factual or CF) explanation e. The feature Γ is therefore excluded from the detailisation store: $DET = DET \setminus \{\Gamma\}$. U is allowed to:

- require further details on another feature of the same explanation remaining in the detailisation store ($\Gamma' \in DET$) (locution *what-details*($\hat{y}$, E[,y'], e, Γ') where $\Gamma' \neq \Gamma$;
- require a CF explanation for an arbitrary CF class $y' \in CFS$ if a factual explanation is being processed (locution *why-not-explain*($\hat{y}$, E, y')) or require a CF explanation for another CF class if a CF explanation is being processed (locution *why-not-explain*($\hat{y}$, E, y''));
- specify how a feature Ψ of the explanation e is defined (locution *what-is*($\hat{y}$, E[,y'], e, Ψ) where $\Psi \in CLAR$;
- disagree with the factual explanation and require an alternative to it, provided that it is the only explanation that E contains (locution *why-alternative*($\hat{y}$, E, e_f)) iff the currently processed explanation is factual (i.e., $e = e^h_f$ and $\nexists\, e_{cf} \in E$);
- require another CF explanation for the same CF class (locution *why-not-alternative*($\hat{y}$, E, y', e_{cf}) iff the explanation currently processed is counterfactual (i.e., $e = e^h_{cf}$);
- accept prediction $\hat{y}$ (locution *accept-u*($\hat{y}$, E));
- reject prediction $\hat{y}$ (locution *reject-u*($\hat{y}$, E)).

(g) *no-elaborate*($\hat{y}$, E[,y'], e, Γ) where $e = e^h_f | e^h_{cf}$: S is unable to provide details on feature $\Gamma \in e$ because either all the available details have already been provided or the details required are not found in the knowledge base of the system. U can respond using one of the following locutions:

- require further details on another feature of the same explanation remaining in the detailisation store ($\Gamma' \in DET$) (locution *what-details*($\hat{y}$, E[,y'], e, Γ') where $\Gamma' \neq \Gamma$;
- require a CF explanation for an arbitrary CF class $y' \in CFS$ if a factual explanation is being processed (locution *why-not-explain*($\hat{y}$, E, y')) or require a CF explanation for another CF class if a CF explanation is being processed (locution *why-not-explain*($\hat{y}$, E, y''));
- specify how a feature Ψ of the explanation e is defined (locution *what-is*($\hat{y}$, E[,y'], e, Ψ) where $\Psi \in CLAR$;
- disagree with the factual explanation and require an alternative to it, provided that it is the only explanation that E contains (locution *why-alternative*($\hat{y}$, E, e_f)) iff the currently processed explanation is factual (i.e., $e = e^h_f$ and $\nexists\, e_{cf} \in E$);
- require another CF explanation for the same CF class (locution *why-not-alternative*($\hat{y}$, E, y', e_{cf}) iff the explanation currently processed is counterfactual (i.e., $e = e^h_{cf}$);
- accept prediction $\hat{y}$ (locution *accept-u*($\hat{y}$, E));
- reject prediction $\hat{y}$ (locution *reject-u*($\hat{y}$, E)).

(h) *clarify*($\hat{y}$, E[,y'], e, Ψ, υ): S provides a definition υ for feature Ψ of the currently processed explanation e. The explanation e can be of any modality: factual or CF, high-level

or low-level. The corresponding feature is then eliminated from the clarification store: $CLAR = CLAR \setminus \{\Psi\}$. U uses one of the following locutions to respond:

- require details on a feature $\Gamma \in e$ remaining in the detailisation store ($\Gamma \in DET$) (locution *what-details*($\hat{y}$, E[,y'], e, Γ));
- require a CF explanation for an arbitrary CF class $y' \in CFS$ if a factual explanation is being processed (locution *why-not-explain*($\hat{y}$, E, y')) or require a CF explanation for another CF class if a CF explanation is being processed (locution *why-not-explain*($\hat{y}$, E, y''));
- specify how another feature Ψ of the explanation e is defined (locution *what-is*($\hat{y}$, E[,y'], e, Ψ') where $\Psi' \in CLAR$;
- require an alternative factual explanation (locution *why-alternative*($\hat{y}$, E, e)) iff $e = e^h_f | e^l_f$;
- require an alternative CF explanation for the same CF class (locution *why-not-alternative* ($\hat{y}$, E, e)) iff $e = e^h_{cf} | e^l_{cf}$;
- accept prediction $\hat{y}$ (locution *accept-u*($\hat{y}$, E));
- reject prediction $\hat{y}$ (locution *reject-u*($\hat{y}$, E)).

(i) *no-clarify*($\hat{y}$, E[,y'], e, Ψ): S is unable to provide a definition of the feature $\Psi \in e$ because the feature is specified incorrectly, or the definition is not found in the system's knowledge base, or the definition has already been provided. U is allowed to respond using one of the following locutions:

- require details on a feature $\Gamma \in e$ remaining in the detailisation store ($\Gamma \in DET$) (locution *what-details*($\hat{y}$, E[,y'], e, Γ));
- require a CF explanation for an arbitrary CF class $y' \in CFS$ if a factual explanation is being processed (locution *why-not-explain*($\hat{y}$, E, y')) or require a CF explanation for another CF class if a CF explanation is being processed (locution *why-not-explain*($\hat{y}$, E, y''));
- specify how another feature Ψ of the explanation e is defined (locution *what-is*($\hat{y}$, E[,y'], e, Ψ') where $\Psi' \in CLAR$;
- require an alternative factual explanation (locution *why-alternative*($\hat{y}$, E, e)) iff $e = e^h_f | e^l_f$;
- require an alternative CF explanation for the same CF class (locution *why-not-alternative*($\hat{y}$, E, e)) iff $e = e^h_{cf} | e^l_{cf}$;
- accept prediction $\hat{y}$ (locution *accept-u*($\hat{y}$, E));
- reject prediction $\hat{y}$ (locution *reject-u*($\hat{y}$, E)).

(j) *alter-f*($\hat{y}$, E, e_f, e'_f): S provides U with a factual explanation e'_f alternative to e_f. The previous (possibly also alternative to the original) piece of factual explanation is removed from the explanation store. The newly generated alternative factual explanation is added to the knowledge store $K = K \cup e'_f$ and the explanation store $E = E \cup e'_f$. The detailisation and clarification stores are populated with the features of the newly generated alternative factual explanation. U responds using one of the following locutions:

- require details on a feature Γ of the offered alternative factual explanation e'_f (locution *what-details(*$\hat{y}$, E, e'_f, Γ*)*);
- require a CF explanation for some CF class y' (locution *why-not-explain*($\hat{y}$, E, y'));
- specify how a feature Ψ of the offered alternative factual explanation e'_f is defined (locution *what-is*($\hat{y}$, E, e'_f, Ψ));
- require another alternative factual explanation (locution *why-alternative*($\hat{y}$, E, e'_f));
- accept prediction $\hat{y}$ (locution *accept-u*($\hat{y}$, E));

- reject prediction $\hat{y}$ (locution *reject-u*($\hat{y}$, E)).

(k) *no-alter-f*($\hat{y}$, E, e_f): S is unable to offer a factual explanation alternative to e_f because there exists no explanation alternative to the factual or all the alternatives have already been offered. U responds using one of the following locutions:

- require details on a feature Γ of the latest offered (either original or alternative) factual explanation (locution *what-details*($\hat{y}$, E, e_f, Γ));
- require a CF explanation for some CF class y' (locution *why-not-explain*($\hat{y}$, E, y'));
- specify how a feature Ψ of the latest offered (either original or alternative) factual explanation is defined (locution *what-is*($\hat{y}$, E, e_f, Ψ));
- accept prediction $\hat{y}$ (locution *accept-u*($\hat{y}$, E));
- reject prediction $\hat{y}$ (locution *reject-u*($\hat{y}$, E)).

(l) *alter-cf*($\hat{y}$, E, y', e_{cf}, e'_{cf}): S provides U with a CF explanation e'_{cf} alternative to e_{cf}. The previous (possibly also alternative to the original) piece of CF explanation is removed from the explanation store. The newly generated alternative CF explanation is added to the knowledge store $K = K \cup e'_{cf}$ and the explanation store $E = E \cup e'_{cf}$. The detailisation and clarification stores are populated with the features of the newly generated alternative CF explanation. U is allowed to respond using one of the following locutions:

- require details on a feature Γ of the offered alternative CF explanation e'_{cf} (locution *what-details*($\hat{y}$, E, y', e'_{cf}, Γ));
- require a CF explanation for some other CF class y'' (locution *why-not-explain*($\hat{y}$, E, y'') iff $y'' \neq y'$);
- specify how a feature Ψ of the offered alternative CF explanation e'_{cf} is defined (locution *what-is*($\hat{y}$, E, y', e'_{cf}, Ψ));
- accept prediction $\hat{y}$ (locution *accept-u*($\hat{y}$, E));
- reject prediction $\hat{y}$ (locution *reject-u*($\hat{y}$, E)).

(m) *no-alter-cf*($\hat{y}$, E, y', e_{cf}): S is unable to offer a CF explanation alternative to e_{cf}. U is allowed to make one of the following actions:

- require details on a feature Γ of the latest offered alternative CF explanation e'_{cf} (locution *what-details*($\hat{y}$, E, y', e'_{cf}, Γ));
- require a CF explanation for some other CF class y'' (locution *why-not-explain*($\hat{y}$, E, y'') iff $y'' \neq y'$);
- specify how a feature Ψ of the latest offered alternative CF explanation e'_{cf} is defined (locution *what-is*($\hat{y}$, E, y', e'_{cf}, Ψ));
- accept prediction $\hat{y}$ (locution *accept-u*($\hat{y}$, E));
- reject prediction $\hat{y}$ (locution *reject-u*($\hat{y}$, E)).

(4) **Termination states.** The dialogue ends when the system generates a concluding locution (either *accept-s*($\hat{y}$, E) or *reject-s*($\hat{y}$, E)) immediately after the end user accepts or rejects the system's prediction, respectively.

An explanatory dialogue is governed in accordance with the aforementioned rules. Table 13 summarises and exemplifies the dialogue protocol outlined above.

Table 13
The set of allowed moves for the participants of an explanatory dialogue game

Locution	Interpretation	Utterance template	Possible response(-s)
		System (S):	
claim($\hat{y}$, E)	S claims prediction $\hat{y}$ to be true	The test instance is of class $\hat{y}$.	• *why-explain*($\hat{y}$, E) • *accept-u*($\hat{y}$, E) • *reject-u*($\hat{y}$, E)
explain-f($\hat{y}$, E, e_f)	S factually explains $\hat{y}$ with e_f	The test instance is of class $\hat{y}$ because $\langle \text{feature}_1 \rangle$ is $\langle \text{term}_1 \rangle$ [and $\langle \text{feature}_2 \rangle$ is $\langle \text{term}_2 \rangle$,...].	• *why-not-explain*($\hat{y}$, E, y') • *what-details*($\hat{y}$, E, e_f, Γ) • *what-is*($\hat{y}$, E, e_f, Ψ) • *why-alternative*($\hat{y}$, E, e_f) • *accept-u*($\hat{y}$, E) • *reject-u*($\hat{y}$, E)
no-explain-f($\hat{y}$, E)	S is unable to factually explain $\hat{y}$	Sorry, I don't have a factual explanation for you.	• *why-not-explain*($\hat{y}$, E, y') • *accept-u*($\hat{y}$, E) • *reject-u*($\hat{y}$, E)
explain-cf($\hat{y}$, E, y', e_{cf})	S counterfactually explains why $\hat{y}$ and not y' with e_{cf}	The test instance would be of class y' if $\langle \text{feature}_1 \rangle$ were $\langle \text{term}_2 \rangle$ [and $\langle \text{feature}_2 \rangle$ were $\langle \text{term}_1 \rangle$,...].	• *what-details*($\hat{y}$, E, y', e_{cf}, Γ) • *why-not-explain*($\hat{y}$, E, y'') iff $y'' \neq y'$ • *what-is*($\hat{y}$, E, y', e_{cf}, Ψ) • *why-not-alternative*($\hat{y}$, E, y', e_{cf}) • *accept-u*($\hat{y}$, E) • *reject-u*($\hat{y}$, E)
no-explain-cf($\hat{y}$, E, y')	S is unable to counterfactually explain why $\hat{y}$ and not y'	Sorry, I don't have a CF explanation for you.	• *why-not-explain*($\hat{y}$, E, y'') iff $y'' \neq y'$ • *why-alternative*($\hat{y}$, E, e_f) iff $\nexists e_{cf} \in E$ • *accept-u*($\hat{y}$, E) • *reject-u*($\hat{y}$, E)
elaborate($\hat{y}$, E[,y'], e, Γ, θ) where $e = e_f^h \vert e_{cf}^h$	S provides requested details θ on feature Γ of a high-level explanation e	I define Γ to be $\langle \text{term} \rangle$, as it ranges from $\langle \min_{\text{term}} \rangle$ to $\langle \max_{\text{term}} \rangle$.	• *what-details*($\hat{y}$, E[,y'], e, Γ') iff $\Gamma' \neq \Gamma$ • *why-not-explain*($\hat{y}$, E, y') iff $e = e_f^h$ or *why-not-explain*($\hat{y}$, E, y'') iff $e = e_{cf}^h$ and $y'' \neq y'$ • *what-is*($\hat{y}$, E[,y'], e, Ψ) • *why-alternative*($\hat{y}$, E,e) iff $e = e_f^h$ • *why-not-alternative*($\hat{y}$, E, y', e) iff $e = e_{cf}^h$ • *accept-u*($\hat{y}$, E) • *reject-u*($\hat{y}$, E)
no-elaborate($\hat{y}$, E[,y'], e, Γ) where $e = e_f^h \vert e_{cf}^h$	S is unable to provide details on feature Γ of a high-level explanation e (e.g., because all the required details have already been provided)	Sorry, I don't have details on Γ.	• *what-details*($\hat{y}$, E[,y'], e, Γ') iff $\Gamma' \neq \Gamma$ • *why-not-explain*($\hat{y}$, E, y') iff $e = e_f^h$ or *why-not-explain*($\hat{y}$, E, y'') iff $e = e_{cf}^h$ and $y'' \neq y'$ • *what-is*($\hat{y}$, E[,y'], e, Ψ) • *why-alternative*($\hat{y}$, E,e) if $e = e_f^h$ • *why-not-alternative*($\hat{y}$, E,e) if $e = e_{cf}^h$ • *accept-u*($\hat{y}$, E) • *reject-u*($\hat{y}$, E)

Table 13
(Continued)

Locution	Interpretation	Utterance template	Possible response(-s)
clarify($\hat{y}$, $E[,y']$, e, Ψ, υ) where $e = e^h_f \vert e^h_{cf} \vert e^l_f \vert e^l_{cf}$	S provides definition υ for feature Ψ making part of explanation e	Ψ is υ.	• *what-details*($\hat{y}$, $E[,y']$, e, Γ) • *why-not-explain*($\hat{y}$, E, y') iff $e = e^h_f \vert e^l_f$ or *why-not-explain*($\hat{y}$, E, y'') iff $e = e^h_{cf} \vert e^l_{cf}$ and $y'' \neq y'$ • *what-is*($\hat{y}$, $E[,y']$, e, Ψ') iff $\Psi' \neq \Psi$ • *why-alternative*($\hat{y}$, E,e) iff $e = e^h_f \vert e^l_f$ • *why-not-alternative*($\hat{y}$, E, y', e) iff $e = e^h_{cf} \vert e^l_{cf}$ • *accept-u*($\hat{y}$, E) • *reject-u*($\hat{y}$, E)
no-clarify($\hat{y}$, $E[,y']$, e, Ψ) where $e = e^h_f \vert e^h_{cf} \vert e^l_f \vert e^l_{cf}$	S is unable to provide a definition for feature $\Psi \in e$ (e.g., because it is absent in the knowledge base or the inquired term is not found in the set of features)	Sorry, I cannot clarify what Ψ is.	• *what-details*($\hat{y}$, $E[,y']$, e, Γ) • *why-not-explain*($\hat{y}$, E, y') iff $e = e^h_f \vert e^l_f$ or *why-not-explain*($\hat{y}$, E, y'') iff $e = e^h_{cf} \vert e^l_{cf}$ and $y'' \neq y'$ • *what-is*($\hat{y}$, $E[,y']$, e, Ψ') where $\Psi' \neq \Psi$ • *why-alternative*($\hat{y}$, E, e) iff $e = e^h_f \vert e^l_f$ • *why-not-alternative*($\hat{y}$, E, e) iff $e = e^h_{cf} \vert e^l_{cf}$ • *accept-u*($\hat{y}$, E) • *reject-u*($\hat{y}$, E)
alter-f($\hat{y}$, E, e_f, e'_f)	S provides a factual explanation e'_f alternative to e_f	The test instance is of class $\hat{y}$ because $\langle \text{feature}_1 \rangle$ is $\langle \text{term}_3 \rangle$ [and $\langle \text{feature}_2 \rangle$ is $\langle \text{term}_4 \rangle$,...].	• *what-details*($\hat{y}$, E, e'_f, Γ) • *why-not-explain*($\hat{y}$, E, y') • *what-is*($\hat{y}$, E, e'_f, Ψ) • *why-alternative*($\hat{y}$, E, e'_f) • *accept-u*($\hat{y}$, E) • *reject-u*($\hat{y}$, E)
no-alter-f($\hat{y}$, E, e_f)	S is unable to provide a factual explanation alternative to e_f	Sorry, I don't have an alternative factual explanation for you.	• *what-details*($\hat{y}$, E, e_f, Γ) • *why-not-explain*($\hat{y}$, E, y') • *what-is*($\hat{y}$, E, e_f, Ψ) • *accept-u*($\hat{y}$, E) • *reject-u*($\hat{y}$, E)
alter-cf($\hat{y}$, E, y', e_{cf}, e'_{cf})	S provides a CF explanation e'_{cf} alternative to e_{cf} for some CF class y'	The test instance would be of class y' if $\langle \text{feature}_1 \rangle$ were $\langle \text{term}_4 \rangle$ [and $\langle \text{feature}_2 \rangle$ were $\langle \text{term}_3 \rangle$,...].	• *what-details*($\hat{y}$, E, y', e'_{cf}, Γ) • *why-not-explain*($\hat{y}$, E, y'') iff $y'' \neq y'$ • *what-is*($\hat{y}$, E, y', e'_{cf}, Ψ) • *accept-u*($\hat{y}$, E) • *reject-u*($\hat{y}$, E)
no-alter-cf($\hat{y}$, E, y', e_{cf})	S is unable to provide a CF explanation alternative to e_{cf}	Sorry, I don't have an alternative CF explanation for you.	• *what-details*($\hat{y}$, E, y', e_{cf}, Γ) • *why-not-explain*($\hat{y}$, E, y'') iff $y'' \neq y'$ • *what-is*($\hat{y}$, E, y', e_{cf}, Ψ) • *accept-u*($\hat{y}$, E) • *reject-u*($\hat{y}$, E)
accept-s($\hat{y}$, E)	S utters the farewell locution, as the user accepted the system's claim	Ok, thank you for your trust in me. Bye!	–

Table 13
(Continued)

Locution	Interpretation	Utterance template	Possible response(-s)
reject-s($\hat{y}$, E)	S utters the farewell locution, as the user rejected the system's claim	Sorry about my poor explanatory capacities. Bye!	–
		User (U):	
why-explain($\hat{y}$, E)	U requests to factually explain prediction $\hat{y}$	Could you explain why you think so?	• *explain-f*($\hat{y}$, E, e_f) • *no-explain-f*($\hat{y}$, E)
why-not-explain($\hat{y}$, E, y')	U requests to counterfactually explain why $\hat{y}$ and not y'	But why not y'?	• *explain-cf*($\hat{y}$, E, y', e_{cf}) • *no-explain-cf*($\hat{y}$, E, y')
what-details($\hat{y}$, E[,y'], e, Γ) where $e = e^h_f \vert e^h_{cf}$	U requests details on a specific feature Γ of a (factual or counterfactual) high-level explanation e ($\Gamma \in e$)	Could you provide me with details on Γ?	• *elaborate*($\hat{y}$, E[,y'], e, Γ, θ) • *no-elaborate*($\hat{y}$, E[,y'], e, Γ)
what-is($\hat{y}$, E[,y'], e, Ψ) where $e = e^h_f \vert e^h_{cf} \vert e^l_f \vert e^l_{cf}$	U requests a definition for a specific feature Ψ making part of (factual or counterfactual, high- or low-level) explanation e ($\Psi \in e$)	What do you mean by Ψ?	• *clarify*($\hat{y}$, E[,y'], e, Ψ, υ) • *no-clarify*($\hat{y}$, E[,y'], e, Ψ)
why-alternative($\hat{y}$, E, e_f)	U disagrees with the offered factual explanation e_f and requires an alternative factual explanation	I do not agree (or, I am not satisfied/convinced) with your (factual) explanation. Could you offer me another one?	• *alter-f*($\hat{y}$, E, e_f, e'_f) • *no-alter-f*($\hat{y}$, E, e_f)
why-not-alternative($\hat{y}$, E, y', e_{cf})	U disagrees with the offered CF explanation e_{cf} and requires an alternative CF explanation for some CF class y'	I do not agree (or, I am not satisfied/convinced) with your (CF) explanation. Could you offer me another one?	• *alter-cf* ($\hat{y}$, E, y', e_{cf}, e'_{cf}) • *no-alter-cf*($\hat{y}$, E, y', e_{cf})
accept-u($\hat{y}$, E)	U accepts all pieces of explanation contained in explanation store E and therefore definitely accepts prediction $\hat{y}$	Ok, I trust (or agree/am satisfied/am convinced) with your prediction.	• *accept-s* ($\hat{y}$, E)
reject-u($\hat{y}$, E)	U rejects (a) piece(-s) of explanation contained in explanation store E and therefore definitely rejects prediction $\hat{y}$	I don't really trust (or am not satisfied/am not convinced/agree with) your prediction and you won't be able to convince me.	• *reject-s* ($\hat{y}$, E)

Appendix C. Explanatory dialogue grammar productions

Recall that an EDG can be formalised by means of a context-free grammar $G = \langle N, T, R, S \rangle$ (see Section 3.3 for details). Outlined below is the set of the generalised dataset-independent production rules (R):

(1) DIALOGUE → CLAIM EXPLANATION TERMINATION
(2) CLAIM → The test instance is of class CLASS.
(3) EXPLANATION → FACT-EXPLANATION (CF-EXPLANATION)* | ϵ

(4) TERMINATION → ACCEPT-U ACCEPT-S | REJECT-U REJECT-S
(5) ACCEPT-U → Okay, I trust your prediction.
(6) ACCEPT-S → Thank you for your trust in me. Bye!
(7) REJECT-U → I don't trust your prediction and you won't convince me.
(8) REJECT-S → Sorry for my poor explanatory capacities. Bye!
(9) FACT-EXPLANATION → WHY-EXPLAIN [EXPLAIN-F | NO-EXPLAIN-F]
(10) WHY-EXPLAIN → Could you explain why you think so?
(11) EXPLAIN-F → SURE INTRO-F [B|b]ecause F-EXPL (and F-EXPL)*. [DETAILISATION | CLARIFICATION | ALTERNATIVE-F | ϵ]
(12) INTRO-F → It is of class CLASS | ϵ
(13) F-EXPL → FEATURE is VALUE
(14) NO-EXPLAIN-F → Sorry, I don't have a factual explanation for you.
(15) SURE → Sure! | ϵ
(16) CF-EXPLANATION → WHY-NOT-EXPLAIN [EXPLAIN-CF | NO-EXPLAIN-CF]
(17) WHY-NOT-EXPLAIN → But why is it not of class CLASS?
(18) EXPLAIN-CF → SURE It would be of class CLASS if CF-EXPL (and CF-EXPL)*. [DETAILISATION | CLARIFICATION | ALTERNATIVE-CF | ϵ]
(19) CF-EXPL → FEATURE were VALUE
(20) NO-EXPLAIN-CF → I don't have an explanation for why it is not of class CLASS.
(21) DETAILISATION → WHAT-DETAILS [ELABORATE | NO-ELABORATE] [DETAILISATION | CLARIFICATION | ALTERNATIVE-F | ALTERNATIVE-CF | ϵ]
(22) WHAT-DETAILS → Could you FURTHER specify how TERM FEATURE is defined?
(23) ELABORATE → Sure! FEATURE is defined to be TERM because it lies in the range RANGE.
(24) NO-ELABORATE → Sorry, I don't any FURTHER details on the requested term. [CLARIFICATION | ALTERNATIVE-F | ALTERNATIVE-CF | ϵ]
(25) FURTHER → further | ϵ
(26) CLARIFICATION → WHAT-IS [CLARIFY | NO-CLARIFY]
(27) WHAT-IS → What do you mean by FEATURE?
(28) CLARIFY → FEATURE is DEFINITION. [DETAILISATION | CLARIFICATION | ALTERNATIVE-F | ALTERNATIVE-CF | ϵ]
(29) NO-CLARIFY → Sorry, I cannot clarify the term FEATURE. [DETAILISATION | ALTERNATIVE-F | ALTERNATIVE-CF | ϵ]
(30) ALTERNATIVE-F → WHY-ALTERNATIVE [EXPLAIN-F | NO-EXPLAIN-F]
(31) ALTERNATIVE-CF → WHY-NOT-ALTERNATIVE [EXPLAIN-CF | NO-EXPLAIN-CF]
(32) WHY-ALTERNATIVE → REQ-ALTERNATIVE-BEG EXPL-TYPE-F REQ-ALTERNATIVE-END
(33) WHY-NOT-ALTERNATIVE → REQ-ALTERNATIVE-BEG EXPL-TYPE-CF REQ-ALTERNATIVE-END
(34) REQ-ALTERNATIVE-BEG → I am not quite satisfied with your
(35) REQ-ALTERNATIVE-END → explanation. Could you offer me another one?
(36) EXPL-TYPE-F → factual | ϵ
(37) EXPL-TYPE-CF → counterfactual | ϵ

Table 14
Aggregated self-reported demographic user data for all the use cases

18–25	14 (26.92%)
26–35	24 (46.16%)
36–45	10 (19.23%)
46–55	3 (5.77%)
56–65	1 (1.92%)

(a) Age

Male	28 (53.85%)
Female	24 (46.15%)

(b) Gender

Doctorate (Ph.D)	20 (38.46%)
Master's (M.A./M.Sc.)	25 (48.08%)
Bachelor's (B.A./B.Sc.)	6 (11.54%)
Prefer not to say	1 (1.92%)

(c) Education

Native speaker	20 (38.46%)
Proficient (C2)	17 (32.69%)
Advanced (C1)	10 (19.23%)
Upper intermediate (B2)	5 (9.62%)

(d) English proficiency

Student	30 (57.69%)
Non-student	22 (42.31%)

(e) Occupation

Appendix D. Further details on human evaluation use cases

This appendix outlines the quantitative results of the human evaluation study. First, we report the demographic data of all the study participants who decided to disclose it. Recall that 60 people participated in the evaluation of the proposed dialogue game. All in all, 52 out of all the 60 (86.67%) study participants disclosed their demographic data. In summary, the overall collection of dialogue transcripts is gender-balanced. In addition, the participants who reported their education level had at least a Bachelor degree. Further, all the subjects had at least the B2 level of English proficiency. Table 14 summarises all the self-reported demographic data collected from all the participants.

Subsequently, we provide the reader with the demographic data of the study participants and the process models grouped by use case. Thus, Section D.1 presents the results for the collection of the basketball dataset-related data. Section D.2 displays the results for the beer style classification explanatory dialogues. Section D.3 highlights the results collected for the thyroid disease classification scenario.

D.1. Basketball player position classification

Fourteen (23.33%) of the 60 collected dialogue transcripts relate to the basketball player position dataset. 12 out of the 14 (85.71%) participants who selected the basketball player position scenario attached their demographic data. In summary, 7 (58.33%) participants who chose this scenario and disclosed the demographic data were males, 5 (41.67%) were females. In addition, all the participants who disclosed their demographic data reported to have at least a Bachelor degree and the C1 level of English proficiency. Table 15 summarises all the self-reported demographic data collected from the participants who selected the basketball player position scenario.

Fig. 10 depicts the process model based on the main building blocks (i.e., claim, explanation, and termination) within the collected explanatory dialogues (see Rule 1 of the EDG, Appendix C, for reference). Thus, 12 out of 14 (85.71%) participants required (at least, factual) explanation(-s) for the given prediction. Further, 11 out of 12 (91.67%) such participants accepted the system's prediction after processing the explanation offered. On the contrary, only one out of the 12 (8.33%) participants rejected

Table 15
Self-reported demographic data of the users who interaction with the classifier trained on the basketball player position dataset

18–25	5 (41.67%)
26–35	6 (50.00%)
36–45	1 (8.33%)

(a) Age

Male	7 (58.33%)
Female	5 (41.67%)

(b) Gender

Doctorate (Ph.D)	2 (16.67%)
Master's (M.A./M.Sc.)	8 (66.66%)
Bachelor's (B.A./B.Sc.)	2 (16.67%)

(c) Education

Native speaker	8 (66.66%)
Proficient (C2)	2 (16.67%)
Advanced (C1)	2 (16.67%)

(d) English proficiency

Student	10 (83.33%)
Non-student	2 (16.67%)

(e) Occupation

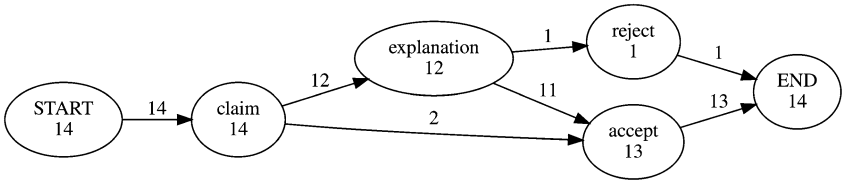


Fig. 10. The process model of the collected basketball player position classification explanatory dialogues based on the main EDG building blocks.

the claim after the explanation was presented. Alternatively, 2 out of 14 (14.29%) participants did not require any explanation for the system's claim. Both of them eventually accepted the system's claim.

As for all the 12 participants who required explanation for the system's claim, 67 explanation-related requests (i.e., those for factual or (alternative) CF explanation, detailisation, and clarification) have been registered. Figure 11 depicts the locution-level process model for the collected explanatory dialogues. Thus, 12 out of the 67 requests (17.91%) were those for factual explanation. In addition, 18 out of the 67 (26.87%) explanation-related requests were those for CF explanation. Further, alternative CF explanations were requested 9 times (13.43%). In addition, 15 out of 67 (22.39%) requests addressed numerical details for the offered linguistic terms whereas only 13 out of 67 requests (19.40%) were clarification requests.

The factual explanation seemed clear and explanatory enough to a half of the participants. Thus, 6 out of 12 (50.00%) study participants who requested a factual explanation did not inquire any further details or clarifications before requesting their first CF explanation. As for the other 6 participants, detailisation requests have been more frequently registered for the factual explanation offered: 7 out of 15 times (46.67%) – 5 (33.33%) times immediately after the factual explanation was offered, 2 (13.33%) times subsequently to the first detailisation request related to the factual explanation. Also, clarification requests are found when processing 6 out of 13 factual explanations (46.15%): once – immediately after it was generated, five times – following detailisation requests. On the other hand, 5 of the 12 (41.67%) participants who requested explanation in the first place were interested in obtaining CF explanations (recall that the 5 participants submitted 18 CF explanation requests altogether). Further, numerical intervals specifying linguistic terms of the corresponding CF explanations were inquired 8 out of the overall

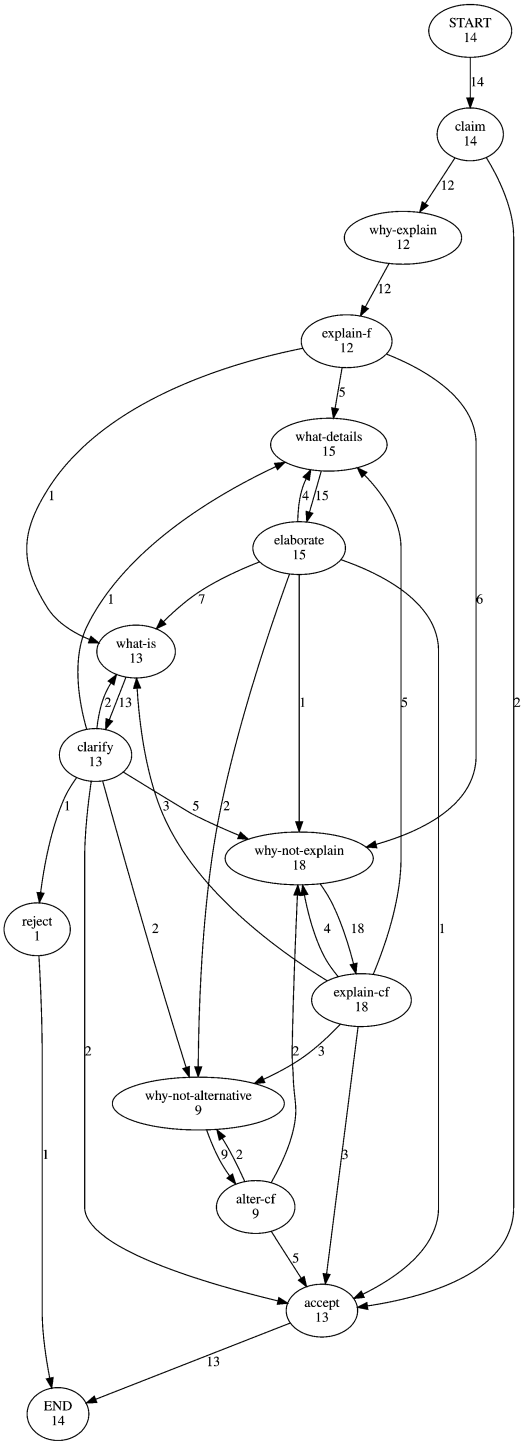


Fig. 11. The full process model of the basketball player position classification explanatory dialogues. For illustrative purposes, pairs of termination nodes, i.e. *{accept-u, accept-s}* and *{reject-u and reject-s}*, are merged into *accept* and *reject*, respectively.

Table 16
Number of times the CF explanations (sorted by rank) were requested by participants (basketball player position classification)

	CF class			
CF rank	*Shooting guard*	*Small forward*	*Power forward*	*Center*
#1	5	6	5	2
#2	3	2	1	1
#3	–	2	–	–

15 times (53.33%), 5 of them submitted as soon as the corresponding CF was offered. In addition, 7 (53.85%) out of all the 13 clarification requests were registered when processing CF explanations, 3 of them – submitted immediately upon receiving the corresponding CF explanation.

Importantly, the locution-level process model (see Fig. 11 for details) shows us the responses to which requests were the most decisive for the study participants to make their final decisions. Recall that 11 out of the 12 participants who required explanation accepted the system's claim. Thus, 3 (27.27%) of the 11 participants accepted the system's claim immediately after a CF explanation had been presented. Further, 5 out of 11 participants (45.46%) found themselves in the position to make the final decision after an alternative CF explanation was displayed. For 2 out of the 11 (18.18%) participants who accepted the claim, the response to their clarification requests triggered their final decision. In addition, 1 out of 11 (9.09%) such participants accepted the system's claim after (s-)he was provided with the details on the inquired feature. Recall that only one subject rejected the claim after having been provided with the explanation. In this case, a response to a clarification request motivated that decision.

Recall that 18 CF explanation requests were registered in the basketball player position classification dialogues. All such CF explanations are those deemed most relevant to the test instance by the system. However, there have as well been registered 9 requests for alternative CF explanations, 7 of them being an alternative to the best ranked CFs.

Table 16 presents numbers of CF explanation requests for each CF class (row "#1") as well those related to second and third best-ranked alternative CF explanations (rows "#2" and "#3", accordingly). Thus, in 7 out of the 18 (38.89%) cases where (the best-ranked) CF explanations were offered, the users did not find them satisfactory. Further, when exposed to 2 out of the 7 (28.57%) second-best ranked CF explanations were offered, the participants required third-best ranked CFs. In particular, both such cases occur when CF explanations were asked for the CF class "Small forward". Importantly, 5 out of all the 9 (55.56%) alternative CF explanations turned out to be crucially decisive from the end user's point of view (i.e., they led to making an immediate decision – in this case, acceptance of the system's claim).

D.2. Beer style classification

Thirty-seven (61.67%) of all the collected dialogue transcripts relate to the beer style classification scenario. All in all, 31 out of the 37 (83.78%) participants who played the beer scenario disclosed their demographic data. In summary, 17 (54.84%) of all the participants who chose this scenario and left their demographic data were males, 14 (45.16%) – females. In addition, all the participants who reported their education level had at least a Bachelor degree and the B2 level of English proficiency. Table 17 summarises all the self-reported demographic data collected from the participants who selected the beer style dataset as the basis of the dialogue game.

Fig. 12 illustrates the process model corresponding to the three main building blocks of the proposed dialogue game. Thus, 36 out of 37 (97.30%) participants required (at least, factual) explanation for the given prediction. Further, 33 out of the 36 (91.67%) participants accepted the system's prediction after

Table 17
Self-reported demographic user data (the beer style classification dataset)

18–25	6 (19.35%)
26–35	14 (45.16%)
36–45	8 (25.81%)
46–55	2 (6.45%)
56–65	1 (3.23%)

(a) Age

Male	17 (54.84%)
Female	14 (45.16%)

(b) Gender

Doctorate (Ph.D)	16 (51.61%)
Master's (M.A./M.Sc.)	12 (38.71%)
Bachelor's (B.A./B.Sc.)	2 (6.45%)
Prefer not to say	1 (3.23%)

(c) Education

Native speaker	9 (29.03%)
Proficient (C2)	11 (35.48%)
Advanced (C1)	8 (25.81%)
Upper intermediate (B2)	3 (9.68%)

(d) English proficiency

Student	14 (45.16%)
Non-student	17 (54.84%)

(e) Occupation

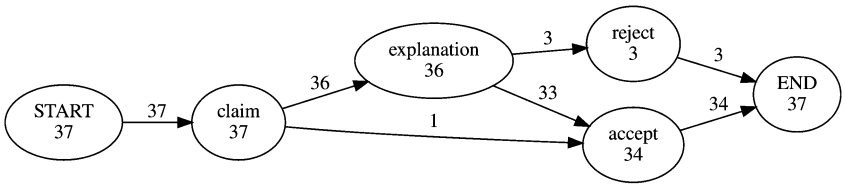


Fig. 12. The process model of the collected beer style classification explanatory dialogues based on the main EDG building blocks.

processing the explanation offered whereas only 3 (8.33%) rejected the system's prediction. In addition, only 1 out of 37 (2.70%) participants did not require any explanation for the system's claim. Eventually, that participant accepted the system's claim.

Figure 13 depicts the locution-level process model for the collected explanatory dialogues. Thus, 235 explanation-related requests (all those covered by the EXPLANATION non-terminal in EDG) were registered from the 36 participants who required explanation for the system's claim. More precisely, 36 out of the 235 (15.32%) requests were those for factual explanation. In addition, 50 out of the 235 (21.28%) explanation-related requests were those for CF explanation. Further, alternative CF explanations were requested 25 times (10.64%). Moreover, 78 out of 235 (33.19%) requests addressed numerical details for the offered linguistic terms whereas 46 out of 235 (19.57%) requests were clarification requests.

It is worth noting that the factual explanation seemed rather unclear to most of the participants. Thus, 31 out of the 36 (86.11%) study participants who requested a factual explanation inquired either further details or clarifications before requesting their first CF explanation. Thus, 52 out of all the 78 detailisation requested registered were concerned with the factual explanation. In 24 (46.15%) cases, numerical intervals for specific features were requested as soon as the factual explanation was presented whereas the other 28 (53.85%) cases of detailisation requests were follow-ups to other (including detailisation) requests. Also, 32 out of 46 (69.57%) clarification requests were found when processing the factual explanation: 7 times (21.88%) – immediately after it was generated, 25 times (78.12%) – following

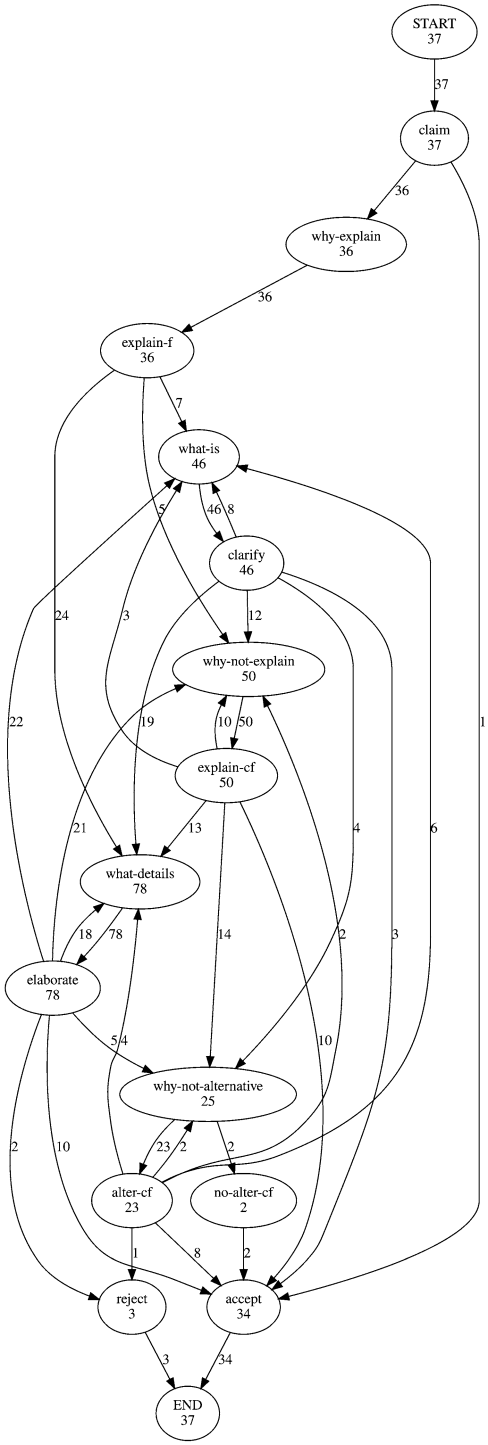


Fig. 13. The full process model of the collected beer style classification explanatory dialogues. For illustrative purposes, pairs of termination nodes, i.e. *{accept-u, accept-s}* and *{reject-u and reject-s}*, are merged into *accept* and *reject*, respectively.

Table 18
Number of times the CF explanations (sorted by rank) were requested by participants (beer style classification)

	CF class						
CF rank	*Lager*	*Pilsner*	*IPA*	*Barleywine*	*Stout*	*Porter*	*BSA*
# 1	12	9	10	5	5	3	6
# 2	7	6	3	1	3	2	1
# 3	1	1	–	–	–	–	–

detailisation or other clarification requests. On the other hand, 29 of the 36 (80.56%) participants who requested explanation in the first place were interested in obtaining CF explanations. Further, numerical intervals specifying linguistic terms of the corresponding CF explanations were inquired 26 out of the overall 78 times (33.33%), a half of them submitted as soon as the corresponding CF was offered. In addition, 14 (30.43%) out of all the 46 clarification requests were registered when processing CF explanations, 3 of them – immediately after the CF explanation was presented. Last but not least, out of the 25 alternative CF explanations requested, 14 (56.00%) were requested immediately after the questioned CF was presented whereas 11 (44.00%) – after detailisation or clarification requests concerning the CF explanation in question or subsequent to other alternative CF requests.

The locution-level process model (see Fig. 13 for details) also shows the responses to which requests were the most decisive for the study participants to make their final decisions in the beer style classification scenario. Thus, 10 out of the 33 participants (30.30%) who inquired an explanation and accepted the system's claim did so immediately after a CF explanation was presented. Further, 8 out of 33 participants (24.24%) found themselves in the position to make the final decision after an alternative CF explanation was displayed. In addition, 2 out of 33 participants (6.06%) accepted the system's prediction despite the fact that the system could not offer the participant an alternative CF upon request. In addition, for 3 out of the 33 (9.09%) participants who accepted the claim, the response to their clarification requests triggered their final decision. Finally, 10 out of 33 (30.30%) such participants accepted the system's claim after (s-)he was provided with the details on the inquired feature. On the other hand, 2 out of 3 (66.67%) participants rejected the claim when offered details on a specific explanation feature whereas 1 out of 3 (33.33%) did so upon receiving an alternative CF explanation.

Recall that 50 CF explanation requests were registered in the beer style classification scenario dialogues. The best ranked CFs (from the system's points of view) were questioned in 23 out of 50 (46.00%) cases, as the participants asked for an alternative CF explanation. Further, in 2 of the 23 (8.70%) such cases, third-best ranked CFs were requested. Table 18 shows the distribution of requests for CF explanation as well as their alternative variants by CF class. Remarkably, 8 out of the 25 (32.00%) alternative CFs turned out to be decisive (led to immediate acceptance of the system's prediction) whereas 1 alternative CF (4.00%) motivated immediate rejection of the system's claim. Finally, 2 out of all the 33 (6.06%) positive decisions made by the participants who requested an explanation were made after the system did not manage to offer an alternative CF explanation.

D.3. Thyroid diagnosis classification

Nine (15.00%) of all the 60 collected dialogue transcripts relate to the thyroid disease dataset. In summary, 4 (44.44%) participants who chose this scenario were males, 5 (55.56%) were females. Similarly to the other classification scenarios, all the participants reported to, at least, have a Bachelor degree and the B2 level of English proficiency. Table 19 summarises all the self-reported demographic data collected from the participants who selected the thyroid disease classification scenario.

Table 19
Self-reported demographic data of the users who interacted with the classifier trained on the thyroid disease dataset

18–25	3 (33.33%)
26–35	4 (44.45%)
36–45	1 (11.11%)
46–55	1 (11.11%)

(a) Age

Male	4 (44.44%)
Female	5 (55.56%)

(b) Gender

Doctorate (Ph.D)	2 (22.22%)
Master's (M.A./M.Sc.)	5 (55.56%)
Bachelor's (B.A./B.Sc.)	2 (22.22%)

(c) Education

Native speaker	3 (33.33%)
Proficient (C2)	4 (44.45%)
Upper intermediate (B2)	2 (22.22%)

(d) English proficiency

Student	6 (66.67%)
Non-student	3 (33.33%)

(e) Occupation

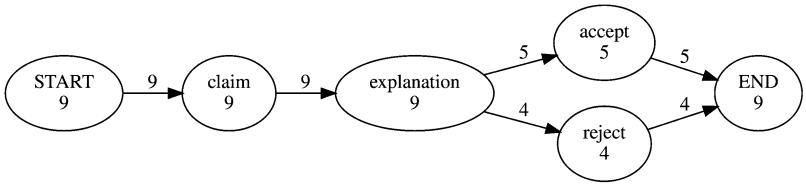


Fig. 14. The process model of the collected thyroid diagnosis classification explanatory dialogues based on the main EDG building blocks.

Fig. 14 illustrates the process model related to the three main building blocks of the proposed model of explanatory dialogue. Thus, all the 9 out of 9 (100.00%) participants required (at least, factual) explanation for the given prediction. Eventually, 5 out of 9 (55.56%) participants accepted the system's claim. On the contrary, 4 out of 9 (44.44%) study participants rejected the system's claim.

As for all the 9 participants who required explanation for the system's claim, 29 explanation-related requests have been registered. Figure 15 depicts the locution-level process model for the corresponding collection of explanatory dialogues. Due to the design of the protocol, 9 out of the 29 requests (31.04%) were those for factual explanation. In addition, 8 out of the 29 (27.59%) explanation-related requests were those for CF explanation. Further, 3 alternative CFs were inquired (10.34% of the explanation-related requests). In addition, 6 out of the 29 (20.69%) requests addressed numerical details for the offered linguistic terms whereas only 3 out of 29 requests (10.34%) were clarification requests.

Out of the nine participants who required (factual) explanation for the system's claim, three (33.33%) requested details for one of the corresponding features that the factual explanation contained. In addition, one participant (11.11%) requested to clarify a term that the factual explanation contained. Besides, one person (11.11%) concluded the dialogue by accepting the system's claim immediately after the factual explanation was displayed whereas four (44.44%) study participants inquired a CF explanation right after processing the factual explanation. Most of the detailisation (4 out of 6, 66.67%) and clarification requests (2 out of 3, 66.67%) addressed the factual explanation. All but one detailisation requests were submitted to the system as soon as the factual explanation was processed whereas one detailisation request followed one of the previously sent detailisation requests. One of the clarification requests was

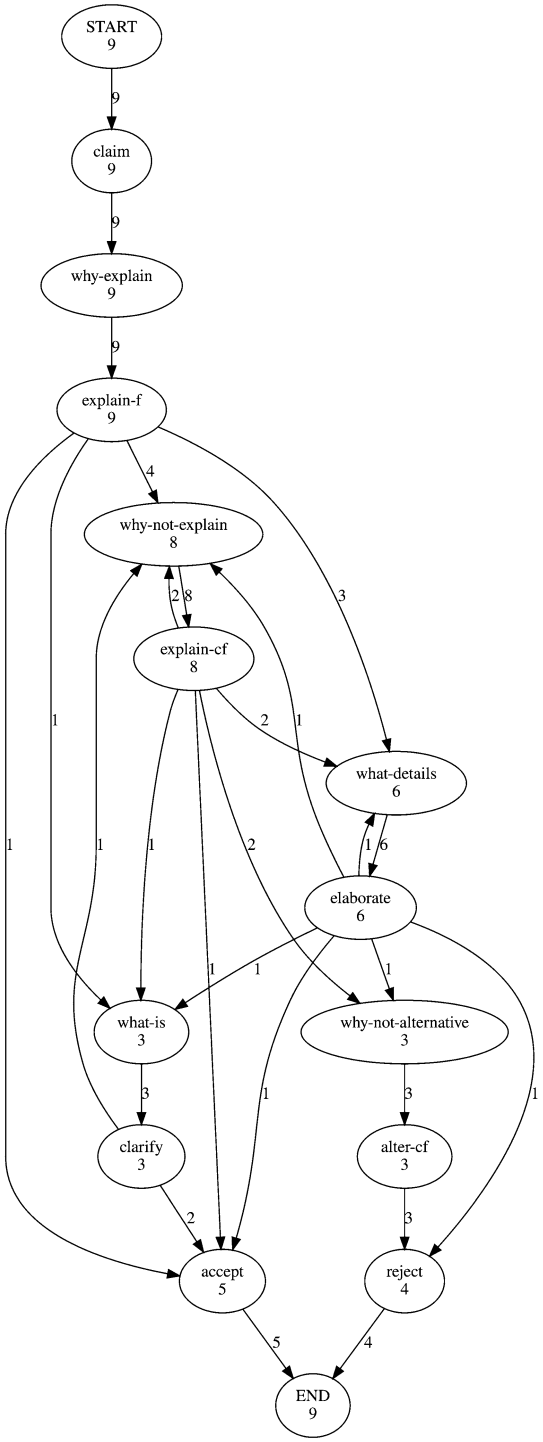


Fig. 15. The full process model of the collected thyroid disease classification explanatory dialogues. For illustrative purposes, pairs of termination nodes, i.e. *{accept-u, accept-s}* and *{reject-u and reject-s}*, are merged into *accept* and *reject*, respectively.

Table 20
Number of times the CF explanations (sorted by rank) were requested by participants (thyroid disease classification)

	CF class		
CF rank	*No hypothyroid*	*Primary hypothyroid*	*Compensatory hypothyroid*
# 1	4	2	2
# 2	2	1	–

sent to the system immediately after the factual explanation was generated whereas the other clarification request followed a detailisation request. Conversely, two (33.33%) detailisation requests were registered for all the CF explanations generated.

The locution-level process model (see Fig. 15 for details) shows the responses to which requests were the most decisive for the study participants to make their final decisions. Out of the 5 participants who accepted the system's claim, one (20.00%) did so immediately after a factual explanation was presented. Similarly, 1 out of 5 (20.00%) accepted the claim in response to a CF explanation offered and a detailisation request each. In addition, 2 (40.00%) participants accepted the system's claim after having received a response to their clarification requests. Out of the 4 participants who rejected the claim, three (75.00%) did so after an alternative CF explanation was offered whereas one (25.00%) was driven by a response to his or her detailisation request.

Finally, recall that 8 CF explanation requests were registered in the thyroid disease classification dialogues. Table 20 presents occurrences of CF explanation requests for each CF class as well those related to alternative CF explanations. Thus, three participants requested an alternative CF explanation for the CF class (two for the class "No hypothyroid" and one – for the class "Primary hypothyroid". Hence, almost a half of the CF explanation requests (3 out of 8, 37.50%) left end users with unsatisfactory responses. Further, all the three such alternative CF explanations turned out to be the final users' dialogue moves before they made their decision (in all the cases the system's claim was eventually rejected).

Acknowledgements

Ilia Stepin is an *FPI* researcher (PRE2019-090153). This work was supported by the Spanish Ministry of Science, Innovation and Universities (grants PID2021-123152OB-C21, TED2021-130295B-C33 and RED2022-134315-T) and the Galician Ministry of Culture, Education, Professional Training and University (grants ED431G2019/04 and ED431C2022/19). These grants were co-funded by the European Regional Development Fund (ERDF/FEDER program). Katarzyna Budzynska was funded in part by POB CyberDS of Warsaw University of Technology within the *Excellence Initiative: Research University* (IDUB) programme under grant 1820/1/Z01/POB3/2021 and in part by VW foundation (VolkswagenStiftung) under grant 98 542. In addition, the authors would like to express their gratitude to Dr. Elena Musi for her invaluable support in gathering experimental data.

References

[1] A. Adadi and M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), *IEEE Access* **6** (2018), 52138–52160. doi:10.1109/ACCESS.2018.2870052.

[2] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J.M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. del Ser, N. Díaz-Rodríguez and F. Herrera, Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence, *Information Fusion* (2023). doi:10.1016/j.inffus.2023.101805.

[3] J.M. Alonso, Teaching explainable artificial intelligence to high school students, *International Journal of Computational Intelligence Systems* **13**(1) (2020), 974–987. doi:10.2991/ijcis.d.200715.003.
[4] A. Arioua, P. Buche and M. Croitoru, Explanatory dialogues with argumentative faculties over inconsistent knowledge bases, *Expert Systems with Applications* **80** (2017), 244–262. doi:10.1016/j.eswa.2017.03.009.
[5] A. Arioua and M. Croitoru, Formalizing explanatory dialogues, in: *International Conference on Scalable Uncertainty Management*, Springer, 2015, pp. 282–297. doi:10.1007/978-3-319-23540-0_19.
[6] F. Behrens, S. Bischoff, P. Ladenburger, J. Rückin, L. Seidel, F. Stolp, M. Vaichenker, A. Ziegler, D. Mottin, F. Aghaei et al., MetaExp: Interactive explanation and exploration of large knowledge graphs, in: *Companion Proceedings of the Web Conference 2018*, 2018, pp. 199–202. doi:10.1145/3184558.3186978.
[7] T.J.M. Bench-Capon, D. Lowes and A.M. McEnery, Argument-based explanation of logic programs, *Knowledge-Based Systems* **4**(3) (1991), 177–183, ISSN 0950-7051. doi:10.1016/0950-7051(91)90007-O.
[8] F. Bex and D. Walton, Combining explanation and argumentation in dialogue, *Argument & Computation* **7**(1) (2016), 55–68. doi:10.3233/AAC-160001.
[9] K. Budzynska, A. Rocci and O. Yaskorska, Financial dialogue games: A protocol for earnings conference calls, in: *Computational Models of Argument (COMMA)*, IOS Press, 2014, pp. 19–30. doi:10.3233/978-1-61499-436-7-19.
[10] R. Calegari, A. Omicini, G. Pisano and G. Sartor, Arg2P: An argumentation framework for explainable intelligent systems, *Journal of Logic and Computation* **32** (2022), 0955. doi:10.1093/logcom/exab089.
[11] R. Calegari, A. Omicini and G. Sartor, Argumentation and logic programming for explainable and ethical AI, in: *Proceedings of the Italian Workshop on Explainable Artificial Intelligence Co-Located with 19th International Conference of the Italian Association for Artificial Intelligence (XAI.it@AIxIA)*, 2020, pp. 55–68.
[12] E. Cambria, L. Malandri, F. Mercorio, M. Mezzanzanica and N. Nobani, A survey on XAI and natural language explanations, *Information Processing & Management* **60**(1) (2023), 103111–103116. doi:10.1016/j.ipm.2022.103111.
[13] G. Castellano, C. Castiello and A.M. Fanelli, The FISDeT software: Application to beer style classification, in: *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2017, pp. 1–6. doi:10.1109/FUZZ-IEEE.2017.8015503.
[14] A. Cawsey, *Explanation and Interaction: The Computer Generation of Explanatory Dialogues*, MIT Press, 1992. doi:10.1007/BF00387398.
[15] F. Cheng, Y. Ming and H. Qu, Dece: Decision explorer with counterfactual explanations for machine learning models, *IEEE Transactions on Visualization and Computer Graphics* **27**(2) (2020), 1438–1447. doi:10.1109/TVCG.2020.3030342.
[16] K. Čyras, A. Rago, E. Albini, P. Baroni and F. Toni, Argumentative XAI: A survey, in: *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI)*, 2021, pp. 4392–4399, Survey Track. doi:10.24963/ijcai.2021/600.
[17] S. Dandl, C. Molnar, M. Binder and B. Bischl, Multi-objective counterfactual explanations, in: *International Conference on Parallel Problem Solving from Nature*, Springer, 2020, pp. 448–469. doi:10.1007/978-3-030-58112-1_31.
[18] G. De Toni, P. Viappiani, B. Lepri and A. Passerini, Generating personalized counterfactual interventions for algorithmic recourse by eliciting user preferences, 2022, arXiv preprint arXiv:2205.13743.
[19] D. Dua and C. Graff, *UCI Machine Learning Repository*, 2017, http://archive.ics.uci.edu/ml.
[20] A. D'Ulizia, F. Ferri and P. Grifoni, A hybrid grammar-based approach to multimodal languages specification, in: *OTM Confederated International Conferences on the Move to Meaningful Internet Systems*, Springer, 2007, pp. 367–376. doi:10.1007/978-3-540-76888-3_59.
[21] D. Engelmann, J. Damasio, A.R. Panisson, V. Mascardi and R.H. Bordini, Argumentation as a method for explainable AI: A systematic literature review, in: *Proceedings of the 17th Iberian Conference on Information Systems and Technologies (CISTI)*, IEEE, 2022, pp. 1–6. doi:10.23919/CISTI54924.2022.9820411.
[22] A. Eshghi, I. Shalyminov and O. Lemon, Bootstrapping incremental dialogue systems from minimal data: The generalisation power of dialogue grammars, in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2220–2230. doi:10.18653/v1/D17-1236.
[23] J. Geertzen, Dialogue act prediction using stochastic context-free grammar induction, in: *Proceedings of the EACL Workshop on Computational Linguistic Aspects of Grammatical Inference*, 2009, pp. 7–15.
[24] A. Ghazimatin, S. Pramanik, R. Saha Roy and G. Weikum, ELIXIR: Learning from user feedback on explanations to improve recommender models, in: *Proceedings of the Web Conference*, 2021, pp. 3850–3860. doi:10.1145/3442381.3449848.
[25] H.P. Grice, Logic and conversation, in: *Syntax and Semantics: Speech Acts*, P. Cole and J.L. Morgan, eds, Academic Press, 1975, pp. 41–58. doi:10.1163/9789004368811_003.
[26] A. Groza, L. Toderean, G.A. Muntean and S.D. Nicoara, Agents that argue and explain classifications of retinal conditions, *Journal of Medical and Biological Engineering* **41**(5) (2021), 730–741. doi:10.1007/s40846-021-00647-7.
[27] R. Guidotti, Counterfactual explanations and how to find them: Literature review and benchmarking, *Data Mining and Knowledge Discovery* (2022), 1–55. doi:10.1007/s10618-022-00831-6.

[28] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri and F. Turini, Factual and counterfactual explanations for black box decision making, *IEEE Intelligent Systems* **34**(6) (2019), 14–23. doi:10.1109/MIS.2019.2957223.
[29] D. Gunning and D. Aha, DARPA's explainable artificial intelligence (XAI) program, *AI magazine* **40**(2) (2019), 44–58. doi:10.1609/aimag.v40i2.2850.
[30] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten, The WEKA data mining software: An update, *SIGKDD Explor. Newsl.* **11**(1) (2009), 10–18, ISSN 1931-0145. doi:10.1145/1656274.1656278.
[31] C.G. Hempel, *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, The Free Press, New York, 1965.
[32] G. Hindelang, Dialogue grammar. A linguistic approach to the analysis of dialogue, *Concepts of Dialogue. Considered from the Perspective of Different Disciplines* (1994), 37–48. doi:10.1515/9783111332062-004.
[33] M. Hirzel, L. Mandel, A. Shinnar, J. Siméon and M. Vaziri, I can parse you: Grammars for dialogs, in: *2nd Summit on Advances in Programming Languages (SNAPL)*, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017. doi:10.4230/LIPIcs.SNAPL.2017.6.
[34] E. Iosif, I. Klasinas, G. Athanasopoulou, E. Palogiannidi, S. Georgiladakis, K. Louka and A. Potamianos, Speech understanding for spoken dialogue systems: From corpus harvesting to grammar rule induction, *Computer Speech & Language* **47** (2018), 272–297. doi:10.1016/j.csl.2017.08.002.
[35] A.-H. Karimi, G. Barthe, B. Balle and I. Valera, Model-agnostic counterfactual explanations for consequential decisions, in: *International Conference on Artificial Intelligence and Statistics, PMLR*, 2020, pp. 895–905.
[36] N.C. Karunatillake, N.R. Jennings, I. Rahwan and P. McBurney, Dialogue games that agents play within a society, *Artificial intelligence* **173** (2009), 935–981.
[37] M.T. Keane, E.M. Kenny, E. Delaney and B. Smyth, If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques, in: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2021, pp. 4466–4474. doi:10.24963/ijcai.2021/609.
[38] I. Klasinas, A. Potamianos, E. Iosif, S. Georgiladakis and G. Mameli, Web data harvesting for speech understanding grammar induction, in: *Interspeech*, 2013, pp. 2733–2737. doi:10.21437/Interspeech.2013-627.
[39] M. Koit, O. Gerassimenko, R. Kasterpalu, A. Raabis and K. Strandson, Towards computer-human interaction in natural language, *International journal of computer applications in technology* **34**(4) (2009), 291–297. doi:10.1504/IJCAT.2009.024082.
[40] S. Kottur, J.M.F. Moura, D. Parikh, D. Batra and M. Rohrbach, CLEVR-dialog: A diagnostic dataset for multi-round reasoning in visual dialog, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 582–595. doi:10.18653/v1/N19-1058.
[41] M. Kuźba and P. Biecek, What would you ask the machine learning model? Identification of user needs for model explanations based on human-model conversations, in: *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2020, pp. 447–459. doi:10.1007/978-3-030-65965-3_30.
[42] C. Labreuche, N. Maudet, W. Ouerdane and S. Parsons, A dialogue game for recommendation with adaptive preference models, in: *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems*, 2015. doi:10.5555/2772879.2773275.
[43] J. Lawrence, M. Snaith, B. Konat, K. Budzynska and C. Reed, Debating technology for dialogical argument: Sensemaking, engagement, and analytics, *ACM Trans. Internet Technol.* **17**(3) (2017). ISSN 1533-5399. doi:10.1145/3007210.
[44] P. McBurney and S. Parsons, Dialogue games for agent argumentation, in: *Argumentation in Artificial Intelligence*, Springer, 2009, pp. 261–280. doi:10.1007/978-0-387-98197-0_13.
[45] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, *Artificial Intelligence* **267** (2019), 1–38. doi:10.1016/j.artint.2018.07.007.
[46] S. Modgil and M. Caminada, Proof theories and algorithms for abstract argumentation frameworks, in: *Argumentation in Artificial Intelligence*, G. Simari and I. Rahwan, eds, Springer US, Boston, MA, 2009, pp. 105–129. doi:10.1007/978-0-387-98197-0_6.
[47] C. Molnar, in: *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*, 2nd edn, Leanpub, 2022, https://christophm.github.io/interpretable-ml-book.
[48] M. Morveli-Espinoza, A. Possebom and C.A. Tacla, A protocol for argumentation-based persuasive negotiation dialogues, in: *Brazilian Conference on Intelligent Systems*, Springer, 2021, pp. 18–32. doi:10.1007/978-3-030-91702-9_2.
[49] R.K. Mothilal, A. Sharma and C. Tan, Explaining machine learning classifiers through diverse counterfactual explanations, in: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 607–617. doi:10.1145/3351095.3372850.
[50] M. Ocana, D. Chapela-Campa, P. Alvarez, N. Hernández, M. Mucientes, J. Fabra, Á. Llamazares, M. Lama, P.A. Revenga, A. Bugarín, M. García-Garrido and J.M. Alonso, Automatic linguistic reporting of customer activity patterns in open malls, *Multimedia Tools and Applications* (2021), 1–27. doi:10.1007/s11042-021-11186-3.

[51] Official Journal of the European Union L119, Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA, Vol. 59, 2016, pp. 89–131, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv.
[52] Official Journal of the European Union L173, Directive 2014/65/EU of the European Parliament and of the Council of 15 May 2014 on markets in financial instruments and amending Directive 2002/92/EC and Directive 2011/61/EU, Vol. 57, 2014, pp. 349–485, https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:32014L0065.
[53] Parliament and Council of the European Union, Proposal for laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts, 2021, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN.
[54] H. Prakken, Coherence and flexibility in dialogue games for argumentation, *Journal of logic and computation* **15**(6) (2005), 1009–1040. doi:10.1093/logcom/exi046.
[55] H. Prakken and R. Ratsma, A top-level model of case-based argumentation for explanation: Formalisation and experiments, *Argument & Computation* **7**(1) (2021), 1–36. doi:10.3233/AAC-210009.
[56] A. Rago, O. Cocarascu, C. Bechlivanidis, D. Lagnado and F. Toni, Argumentative explanations for interactive recommendations, *Artificial Intelligence* **296** (2021), 103506. doi:10.1016/j.artint.2021.103506.
[57] M. Ravi, A. Negi and S. Chitnis, in: *A Comparative Review of Expert Systems, Recommender Systems, and Explainable AI, in: Proceedings of the IEEE 7th International Conference for Convergence in Technology (I2CT)*, 2022, pp. 1–8. doi:10.1109/I2CT54291.2022.9824265.
[58] J.J. Robinson, Diagram: A grammar for dialogues, *Communications of the Association for Computing Machinery* **25**(1) (1982), 27–47. doi:10.1145/358315.358387.
[59] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nature Machine Intelligence* **1**(5) (2019), 206–215. doi:10.1038/s42256-019-0048-x.
[60] C. Russell, Efficient search for diverse coherent explanations, in: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 20–28. doi:10.1145/3287560.3287569.
[61] I. Sassoon, N. Kökciyan, E. Sklar and S. Parsons, Explainable argumentation for wellness consultation, in: *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, Springer, 2019, pp. 186–202. doi:10.1007/978-3-030-30391-4_11.
[62] M. Schleich, Z. Geng, Y. Zhang and D. Suciu, GeCo: Quality counterfactual explanations in real time, *Proc. VLDB Endow.* **14**(9) (2021), 1681–1693. doi:10.14778/3461535.3461555.
[63] U. Schmid and B. Wrede, What is missing in XAI so far?, *KI-Künstliche Intelligenz* **36** (2022), 303–315. doi:10.1007/s13218-022-00786-2.
[64] F. Sebastian, From Speech Act Theory to Dialog: Dialog Grammar, *The Routledge Handbook of Language and Dialogue* (2017), 162–173. doi:10.4324/9781315750583.
[65] A.T.Q. Shaheen and J. Bowles, Dialogue games for explaining medication choices, in: *Procedings of the 4th International Joint Conference on Rules and Reasoning*, Vol. 97, Springer Nature, 2020. doi:10.1007/978-3-030-57977-7_7.
[66] Z. Shams, D.V. Marina, O. Nir and P. Julian, Normative practical reasoning via argumentation and dialogue, in: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, 2016, pp. 1244–1250. doi:10.17863/CAM.58863.
[67] X. Shao, T. Rienstra, M. Thimm and K. Kersting, Towards understanding and arguing with classifiers: Recent progress, *Datenbank-Spektrum* **20**(2) (2020), 171–180. doi:10.1007/s13222-020-00351-x.
[68] H. Shi, R.J. Ross, T. Tenbrink and J. Bateman, Modelling illocutionary structure: Combining empirical studies with formal model analysis, in: *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, 2010, pp. 340–353. doi:10.1007/978-3-642-12116-6_28.
[69] E.I. Sklar and M.Q. Azhar, Explanation through argumentation, in: *Proceedings of the 6th International Conference on Human-Agent Interaction*, 2018, pp. 277–285. doi:10.1145/3284432.3284470.
[70] K. Sokol and P. Flach, One explanation does not fit all: The promise of interactive explanations for machine learning transparency, KI – Künstliche Intelligenz, 2020. ISSN 1610-1987. doi:10.1007/s13218-020-00637-y.
[71] I. Stepin, J.M. Alonso, A. Catala and M. Pereira-Fariña, A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence, *IEEE Access* **9** (2021), 11974–12001, ISSN 2169-3536. doi:10.1109/ACCESS.2021.3051315.
[72] I. Stepin, J.M. Alonso-Moral, A. Catala and M. Pereira-Fariña, An empirical study on how humans appreciate automated counterfactual explanations which embrace imprecise information, *Information Sciences* **618** (2022), 379–399, ISSN 0020-0255. doi:10.1016/j.ins.2022.10.098.
[73] M. Suffian, P. Graziani, J.M. Alonso and A. Bogliolo, FCE: Feedback based counterfactual explanations for explainable AI, *IEEE Access* **10** (2022), 72363–72372. doi:10.1109/ACCESS.2022.3189432.

[74] W.R. Swartout and S.W. Smoliar, On making expert systems more like experts, *Expert Systems* **4**(3) (1987), 196–208. doi:10.1111/j.1468-0394.1987.tb00143.x.
[75] B. Ustun, A. Spangher and Y. Liu, Actionable recourse in linear classification, in: *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT)*, 2019, pp. 10–19. doi:10.1145/3287560.3287566.
[76] W. Van Der Aalst, *Process Mining: Data Science in Action*, Vol. 2, Springer, 2016.
[77] A. Vassiliades, N. Bassiliades and T. Patkos, Argumentation and explainable artificial intelligence: A survey, *The Knowledge Engineering Review* **36** (2021). doi:10.1017/S0269888921000011.
[78] S. Verma, J. Dickerson and K. Hines, Counterfactual explanations for machine learning: A review, in: *Proceedings of the Machine Learning Retrospectives, Surveys & Meta-Analyses (ML-RSA) Workshop at the Conference on Neural Information Processing Systems (NeurIPS)*, 2020.
[79] J. Waa, M. Robeer, J. Diggelen, M. Brinkhuis and M. Neerincx, Contrastive explanations with local foil trees, in: *Proceedings of the International Conference on Machine Learning (ICML) Workshop on Human Interpretability (WHI) in Machine Learning*, 2018.
[80] S. Wachter, B. Mittelstadt and C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, *Harvard Journal of Law & Technology* **31** (2018), 841–887.
[81] D. Walton, Dialogical models of explanation, in: *Proceedings of the Conference on Explanation-Aware Computing (ExaCt) Workshop*, 2007, pp. 1–9.
[82] D. Walton and E.C. Krabbe, *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*, SUNY Press, 1995.
[83] D. Wang, Q. Yang, A. Abdul and B.Y. Lim, Designing theory-driven user-centric explainable AI, in: *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI'19*, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–15. doi:10.1145/3290605.3300831.
[84] K. Weitz, L. Vanderlyn, N.T. Vu and E. André, "It's our fault!": Insights into users' understanding and interaction with an explanatory collaborative dialog system, in: *Proceedings of the 25th Conference on Computational Natural Language Learning, Association for Computational Linguistics*, 2021, pp. 1–16. doi:10.18653/v1/2021.conll-1.1.
[85] T. Wu, M.T. Ribeiro, J. Heer and D. Weld, Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models, in: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, Online, 2021, pp. 6707–6723. doi:10.18653/v1/2021.acl-long.523.
[86] L.A. Zadeh, Linguistic variables, approximate reasoning and dispositions, *Medical Informatics* **8**(3) (1983), 173–186. doi:10.3109/14639238309016081.
[87] D. Zhang, S. Mishra, E. Brynjolfsson, J. Etchemendy, D. Ganguli, B. Grosz, T. Lyons, J. Manyika, J.C. Niebles, M. Sellitto et al., *The AI Index 2022 Annual Report, AI Index Steering Committee, Stanford Institute for Human-Centered AI*, Stanford University, 2022.