Information Processing and Management 60 (2023) 103111

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

A survey on XAI and natural language explanations

Erik Cambria [e], Lorenzo Malandri [a,c], Fabio Mercorio [a,c], Mario Mezzanzanica [a,c], Navid Nobani [b,d,*]

[a] Department of Statistics and Quantitative Methods, University of Milan-Bicocca, Via Bicocca degli Arcimboldi, 8, Milan, 20126, Italy
[b] Department of Informatics, Systems & Communication, University of Milan-Bicocca, Viale Sarca, 336, Milan, 20126, Italy
[c] CRISP Research Centre, University of Milan-Bicocca, Via Bicocca degli Arcimboldi, 8, Milan, 20126, Italy
[d] Digital Attitude S.r.l., Via Giacomo Mellerio, 3, Milan, 20123, Italy
[e] School of Computer Science and Engineering, Nanyang Technological University, 50 Nanyang Avenue, 639798, Singapore

ARTICLE INFO

Keywords:
Explainable AI
Natural language explanations
Presentation methods

ABSTRACT

The field of explainable artificial intelligence (XAI) is gaining increasing importance in recent years. As a consequence, several surveys have been published to explore the current state of the art on this topic. One aspect that seems to be overlooked by these works is the applied presentation methods and, specifically, the role of natural language in generating the final explanations. This survey reviews 70 XAI papers published between 2006 and 2021 and evaluates their readiness with respect to natural language explanations. Thus, together with a set of hierarchical criteria, we define a multi-criteria decision-making model. Finally, we conclude that only a handful of recent XAI works either considered natural language explanations to approach final users (see, e.g.,(Bennetot et al., 2021)) or implemented a method capable of generating such explanations.

1. Introduction and motivation

The need for eXplainable AI (XAI) systems is growing as modern Machine Learning (ML) algorithms, particularly "deep learning" ones, are becoming increasingly powerful yet so complex that is difficult to understand their behaviour and why certain results were achieved or some mistakes were made. However, understanding the behaviour of those models is as relevant as their performances, because it allows users to develop appropriate trust and reliance (Hoffman, Klein and Mueller, 2018). The goal of eXplainable AI is to render the behaviour of black-box models more understandable, accountable and transparent to humans (Burkart & Huber, 2021). This goal can be achieved either by targeting the general decision-making process of a model (Caruana, Lundberg, Ribeiro, Nori, & Jenkins, 2020) or by providing insights about a specific outcome (Cambria, Liu, Decherchi, Xing, & Kwok, 2022; Ehsan, Tambwekar, Chan, Harrison, & Riedl, 2019; Hohman, Srinivasan, & Drucker, 2019; Przybyła & Soto, 2021).

Despite the prominence of XAI methods in recent AI literature and the ever-widening range of their application domains, the attention paid to the *"last mile"* of the XAI-based systems, i.e., the presentation of explanations to end-users is still in a growing phase. This could lead to solutions that, while potent and practical from a technical point of view, cannot be directly utilized by non-expert or non-technical users, defying the principal objective of an XAI system (Miller, 2019). According to the most recent survey on XAI for machine learning models (Burkart & Huber, 2021), the communication type of an XAI system can be classified

* Corresponding author at: Department of Informatics, Systems & Communication, University of Milan-Bicocca, Viale Sarca, 336, Milan, 20126, Italy.
E-mail addresses: cambria@ntu.edu.sg (E. Cambria), lorenzo.malandri@unimib.it (L. Malandri), fabio.mercorio@unimib.it (F. Mercorio), mario.mezzanzanica@unimib.it (M. Mezzanzanica), navid.nobani@unimib.it (N. Nobani).
URLs: https://sentic.net (E. Cambria), https://www.crisp-org.it (M. Mezzanzanica).

https://doi.org/10.1016/j.ipm.2022.103111
Received 26 January 2022; Received in revised form 14 July 2022; Accepted 29 September 2022
Available online 25 October 2022

into *textual*, *graphics*, and *multimedia* descriptions. While the former uses explanations in a text form and the second is a visual one, the latter combines different types of content: text, graphics, reports, images, audio, video, animation, etc.

Considering the works that the authors are aware of, textual explanations can be expressed by rules (Guidotti et al., 2018), codes (Kitzelmann, Schmid, Olsson, & Kaelbling, 2006; Mao, Gan, Kohli, Tenenbaum, & Wu, 2018) or natural language explanations (Mariotti, Alonso, & Gatt, 2020) and dialogues (Jentzsch, Höhn, & Hochgeschwender, 2019). The explanations made through natural language, as pointed out by Mariotti et al. (2020), are the key component of future intelligent interactive agents, given their ability to offer interpretability to people with diverse backgrounds and to better mimic humans, which usually explain their decisions verbally (Burkart & Huber, 2021; Cambria, Hussain, Havasi, & Eckl, 2009). Moreover, in Gkatzia, Lemon, and Rieser (2016), the authors mention that the use of natural language improves decision-making under uncertainty compared to graphical-based presentation methods.

In Sokol and Flach (2018), the authors argue that natural language explanations are suitable for a lay audience given that their interaction mode gives the process a natural feel, while Chaves and Gerosa (2020) and De Gennaro, Krumhuber, and Lucas (2020) make a step further by observing that natural language presentation increases the trustworthiness of the explanations and help in garner user acceptance. These explanations are more efficient (Alonso, Ramos-Soto, Reiter, & van Deemter, 2017), more insightful for specific cases concerning the visual methods (Park et al., 2018) and target a broader range of users. Mariotti et al. (2020) argue that these explanations leverage the common language with the user, profiting from the mental concepts which are already established in the human language (Reiter & Dale, 2000). Krahmer and Theune (2010) argue that generating a text that mimics how humans use natural language to explain, describe, or inform is not a straightforward task, neither to choose the message communicated nor to transform it to natural language.

It is worth mentioning that the majority of works that use natural language presentation methods (i.e., a small part of the XAI works) utilize primary forms of natural language generation (NLG) techniques (Li, Pan, Wang, Yang, & Cambria, 2018) like mail-merge (template-filling) which, while effective and easy to control, as Mariotti et al. (2020) point out, some times produce outputs that are non-natural due to their static nature. The minimal attention paid to the presentation techniques by the current state-of-the-art methods and XAI surveys, makes the process of the XAI method selection by researchers and practitioners time-consuming and error-prone, as the XAI literature lacks a consolidated study on the presentation methods and the way an XAI method should be chosen based on its presentation power. For instance (Burkart & Huber, 2021) allocate a short paragraph to what they call *communication* or Vilone and Longo (2020, 2021) that only briefly mention textual explanations understandability of the explanations for laypeople.

Motivating example. To clarify the matter, let us consider an imaginary conversation between a user (👤) and the system (⚙) that has been trained and deployed to classify million online job ads over a standard occupation taxonomy (Giabelli, Malandri, Mercorio, Mezzanzanica, & Seveso, 2020).

👤: Why have you classified this job ad as *"Data engineer"*?
⚙: I saw *"Data"* and *"Architect"* words.
👤: What is the most important word for classifying this job ad?
⚙: *"Architect"*
👤: Which word you didn't consider for the classification?
⚙: *"Senior"*
👤: What if the input was *"Senior Python data architect"*?
⚙: Then, the class would be *"Data Scientist"*.
👤: Why, instead of classifying it as *"Software Engineer"*, you have chosen *"Data engineer"*?
👤: I never saw titles containing the word *"architect"* being labelled as *"Software engineer"* in the training data.
👤: Why you ignored the word *"senior"*?
⚙: I consider this word as *"too general"*.
👤: Answer me what "too general" means by considering me a data scientist.
⚙: The TF-IDF score of the word "senior"- calculated based on the training data- is distant from the score of the rest of the terms.
👤: What if I tell you that I'm your developer?
⚙: In that case, I would say that I took that decision because the word senior is on the list of stopwords.
👤: Thanks! That's it.
⚙: Before you go... I've noticed that while the word *"senior"* is on the stopword list, the term *"experienced"* which has a high similarity to this word, is not.

To our knowledge, neither in the academic nor in the business world, some methods or frameworks can satisfy all the points mentioned in the above imaginary conversation. Such a system should be able to directly and dynamically interact with the user through natural language, access the black-box, training data and the pipeline, probe them autonomously, identify and act upon users' knowledge, and warn them about possible abnormalities.

Our survey aims to assess the readiness of the current state of the art towards such a solution by mapping them into a collaborative roadmap as shown in Section 2.1, which can provide guidance to academics and practitioners in the study and choice of XAI methods.

We opt for a collaborative roadmap because the research on XAI is growing at a fast pace, both in terms of relevance and the number of methods proposed. According to Vilone and Longo (2021), 250 relevant methods were proposed only between 2016 and 2020, more than four times the ones proposed in the previous five years. Therefore, a static survey would soon become obsolete, losing part of its usefulness.

Contribution. The contribution of this work is two-fold:

1. We provide a survey of presentation techniques used in XAI systems proposing a roadmap covering the whole process of explanation generation, from the black-box to the end-user. Despite the availability of several XAI surveys in the literature (Biran & Cotton, 2017; Burkart & Huber, 2021; Guidotti et al., 2018; Vilone & Longo, 2021), this is the first one that focuses on the presentation methods.
2. Given the short life of classic survey studies, we propose (and make available to the reader) a multi-criteria-decision-making model to allow readers to select the most suited XAI work based on her needs.[1]

2. Does XAI need natural language explanations?

As was mentioned in the introduction section, the literature report various advantages of using natural language methods in explanation creation.

As mentioned in the introduction section, using natural language methods in explanation creation has various advantages like higher efficiency (See Alonso et al. (2017)) and coverage (in terms of audience) (See Sokol and Flach (2018)). Another element is the power of these presentation techniques to convey social cues (Chaves & Gerosa, 2020), interacting with users and reinforcing their trust in the information system as a whole which goes beyond the black-box model and its underneath data (De Gennaro et al., 2020). Such potential benefits can be achieved when XAI solutions, starting from static, one-directional messages, go towards dialogues that directly engage the end-user in the explanation process by offering rich and personalized interactions that mimic how humans explain their decisions.

2.1. A roadmap for selecting XAI-based systems

The definition of XAI is discipline-dependent (Doran, Schulz, & Besold, 2017). Fields close to social and cognitive sciences tend, when defining explanations, to focus on the problem of providing to the end user sufficient information to establish causation (Lipton, 1990; Miller, 2021) while on the other hand, researchers studying human–computer interactions focus on the interactivity, information transition flow and the effectiveness of explanations (Holzinger, Langs, Denk, Zatloukal, & Müller, 2019; Raman et al., 2013). In this paper, we rely on the definition provided by Ribeiro, Singh, and Guestrin (2016), which relates explanations to ML systems and their components (i.e., independent and dependent variables) and is general regarding the study domain: *"textual or visual artefacts that provide a qualitative understanding of the relationship between the instance's components (e.g., words in a text, patches in an image) and the model's prediction"*.

Paper selection criteria. We reviewed 70 XAI papers that make use of natural language. Only those which met these criteria have been included:

- for journal papers, to be either *Q1* or *Q2* of SCImago journal ranking in any computer science-related topics in the year of publication;
- for conference papers, to be classified as A/B for *all* those rankings: (i) CORE Conference Rating, (ii) LiveSHINE, and (iii) Microsoft Academic.

Google Scholar was searched for papers from 2006 to 2021. We identified search terms as combinations of XAI and NLG set of terms and further extended them, identifying new keywords from those papers. The final term sets are XAI (XAI, Explainable Artificial Intelligence, Interpretable AI, Interpretable Artificial Intelligence, Interpretable Machine Learning, IML) and natural language (Natural Language Generation, Natural Language, NLG, verbalization, text).

Finally, we designed a roadmap for the selection of XAI systems studied based on their characteristics. This roadmap consists of three layers, namely, *Context definition*, *Explanation generation* and *Message generation*, in a way that the input of each layer is the output of the previous one. Sections 3 to 5 describe these layers in more detail. A similar approach has been proposed in Hall et al. (2019), that summarized XAI methods by drawing a list of predefined characteristics. However, our proposed roadmap is different from the mentioned work since (i) it uses categorical measures with simple graphics which are faster both for insertion of new papers and interpretation of method comparison (in contrast to textual descriptions used in the above-mentioned work) (ii) it can be used to compute a rank for each entry based on the relative importance of the features based on the demand of the user. In the following three sections we detail each layer of the roadmap based on the literature review mentioned before.

3. Context definition

In our roadmap, the context indicates a layer of information that identifies how the explainer targets the black-box and the need of the end-user (who should use the generated explanations). The context information is used as the input of the explanation generation layer. The building blocks are described below.

[1] A GitHub repository with the ranking model in a machine-readable structure will be provided in case of paper acceptance.

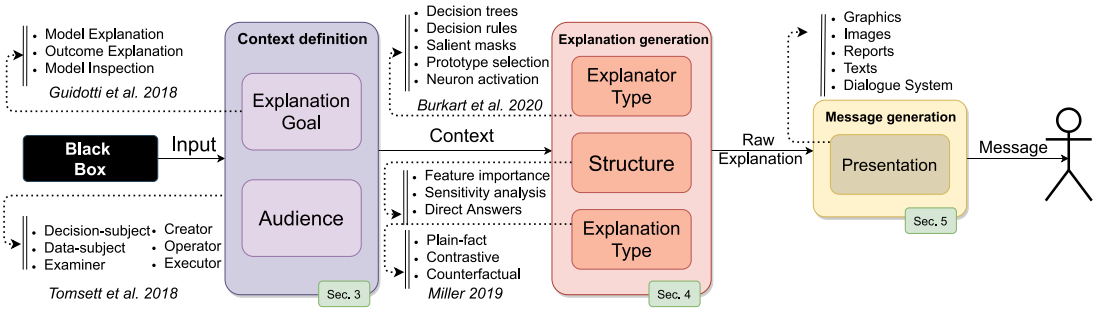

Fig. 1. A roadmap for selecting XAI systems that make use of natural language explanations.

3.1. Explanation goal

In their survey, Guidotti, Monreale, Ruggieri, Turini et al. (2018) divide XAI problems into two broad categories of black-box explanation problems and transparent box design, with the first one further has been categorized into the following three sub-categories:

Model Explanation: Explanations are made through the generation of an interpretable model that tries to mimic the black-box, i.e., generating the same output. If such an interpretable model is successful, generating outputs that are similar to those created by the original model, or in other words, have high fidelity to it, such model can be used as a proxy to understand the general decision-making process of the black-box. By doing so, we can claim that we globally explained the initial black-box.

Outcome explanation: In this case, unlike the model explanations, we are not trying to explain the black-box as a whole; instead, given the record, we want to explain its output as a local explanation. Using such a method means we would not comprehend any more the entire mechanism of the black-box but only a specific outcome of it.

Model inspection: While the outcome of the previous two methods is a model (an interpretable model) which is able to mimic the black-box behaviour (globally in the first and locally in the second case), model inspection, on the other hand, consists of the techniques that instead of generating an interpretable model, provides a visual or textual representation of the model's internal mechanism.

3.2. Audience

XAI studies can be divided into two main groups based on how they address the target users. The first group that makes most works are studies that do not mention the target or audience altogether for their proposed solutions. Almost in all cases, it means that they generate explanations that target technical users who are able to interpret the complex/technical explanations (Goldstein, Kapelner, Bleich, & Pitkin, 2015; Kim, Shah and Doshi-Velez, 2015; Sturm, Lapuschkin, Samek, & Müller, 2016). The second category includes works in which authors mention a general division among different types of audiences (i.e., dividing them into technical and non-technical users), target a specific group of audience or, in rare cases, propose solutions that have a certain level of customization (Alonso & Bugarín, 2019; Hohman et al., 2019).

Tomsett, Braines, Harborne, Preece, and Chakraborty (2018) defines six types of agents – direct and indirect users of an XAI system – in their proposed ecosystem:

- Creator: Agents who create the system, divided into owners and implementors subgroups.
- Operator: Agents that directly interact with the machine.
- Executor: Agents that make decisions based on the output of the AI system.
- Decision subject: Agents that are affected by the decisions.
- Data subject: Agents whose data is used in the targeting of the model.
- Examiners: Agents that audit or investigate the machine.

Similar categories are introduced by previous research, for instance, Bhatt et al. (2020) divides what they called *stakeholders* of explainability to Executive, machine learning engineers, end-users and other stakeholders and Langer et al. (2021), dividing such stakeholders into five groups of users, (system) developers, affected parties, deployers, and regulators.

Several researchers emphasize the importance of providing explanations that are adequate for the audience, fostering interdisciplinary collaboration to maximize the effectiveness of XAI methods in their context of application (Johs, Agosto, & Weber, 2020; Payrovnaziri et al., 2020; Xu et al., 2019). In line with those arguments, we believe that to convey the desired message to the target user successfully and, at the same time, stimulate trust in her, it is necessary to generate explanations tailored to that specific user both in terms of content and form. In our roadmap we use **End users**, **Developers** and **Decision-makers**, as the mostly addressed targets in the literature.

4. Explanation generation

The need for XAI methods is expressed in different forms in the literature; with objectives that sometimes are quite different from each other. Here we mention some works which try to answer the question *What is the necessity of explanations?*:

- Identification of bias & improving fairness (e.g., Guidotti, Monreale, Ruggieri, Turini et al., 2018; Ribera & Lapedriza, 2019; Sokol & Flach, 2020; Wang, Yang, Abdul, & Lim, 2019)
- Trust in the AI systems and algorithmic decision-making processes (e.g., Abdul, Vermeulen, Wang, Lim, & Kankanhalli, 2018; Guidotti, Monreale, Ruggieri, Turini et al., 2018; Hoffman, Mueller, Klein and Litman, 2018; Lucic, Haned, & de Rijke, 2020; Sokol & Flach, 2020)
- Having better control of the AI systems (e.g., Abdul et al., 2018; Wang et al., 2019)
- Debugging and improvement of black-box models (e.g., Kenny, Ford, Quinn, & Keane, 2021; Kulesza, Stumpf, Burnett, & Kwan, 2012; Mittelstadt, Russell, & Wachter, 2019; Sokol & Flach, 2020; Wang et al., 2019)
- Ethical issues (e.g., Apicella, Isgrò, Prevete, & Tamburrini, 2019; Muller, Mayrhofer, Van Veen, & Holzinger, 2021; Stöger, Schneeberger, & Holzinger, 2021)
- Legal issues (e.g., Apicella et al., 2019; Mittelstadt et al., 2019; Ribera & Lapedriza, 2019)
- Improving Transparency (e.g., Abdul et al., 2018; Apicella et al., 2019; Werner, 2020)

Observing from a different angle, Lecue, 2020 identifies the application of explanations in the following fields of AI:

- Machine Learning (except neural networks) (e.g., supervised learning)
- Artificial (Deep) Neural Networks
- Computer Vision
- Constraint Satisfaction and Search (e.g., Conflict resolutions)
- Game Theory (e.g., Zero-sum games)
- Uncertainty in AI (e.g., Probabilistic Graphical Models)
- Robotics (e.g., Information processing)
- Distributed AI (e.g., Multi-Agent Systems)
- Automated Planning and Scheduling (e.g., Unmanned vehicles)
- Natural Language Processing (e.g., Question answering)

Due to its importance, here we briefly describe three types of cognitive processes used in explanations, as outlined by Miller (2019):*Causal connection* or inferring explanation based on the observations and the prior knowledge, *Causal selection* or selecting the inferred explanations and finally, *explanation evaluation* or evaluating the quality of the explanations by the explainee (see e.g., Holzinger, Carrington, & Müller, 2020). He further argues that the ideal outcome of the "explanation evaluation" phase, which is the *best evaluation* is not equivalent to choosing the most likely or the most accurate case, since what is perceived as the best explanation by the explainee is not based on the probability with which the explanation occurs but its pragmatic influence, e.g., usefulness and relevance (see McClure (2002)).

4.1. Explanator type

Regardless of the explanation goal one pursues, there are a variety of explanators (i.e., *"part of the AI system which generates explanation artefacts"* (Hall et al., 2019)). Choosing a model depends on several characteristics of the system like the type of input data and query, accessibility of the black-box, its cost and finally, the context of the explanation as described in Section 3. Burkart and Huber (2021) did a thorough job identifying the principal explanator types. The most used types are: **Decision Trees** (DT), **Decision Rules** (DR), **Salient Masks** (SM) and **Feature Inspection** (FIN). Here, we briefly describe the most used models in the literature these methods:

- **Decision tree (DT)**: Being one of the most-used techniques. Decision trees offer both global and local explanations.
- **Decision rules (DR)**: Decision rules describe the inner mechanism of black-box models by extracting such rules through various methods. Though it is not technically generated through NLG techniques, decision rules often offer explanations that are easy to understand and interpret by a wide range of users.
- **Salient mask (SM)**: Mainly used with image data, salient (or saliency) masks cover certain parts of the input to emphasize the segments used for generating the output.

4.2. Structure

The general form through which the explainer formulates the explanations can be categorized into the following groups:

- **Feature importance (FI)**: Providing a complete or limited set of features with their contribution to the final results. As we discussed in Section 3, such *"final results"* could refer to a single record's output or the general decision-making logic of the black-box model. While used on a large scale, feature importance is not able to demonstrate the root cause of a phenomenon or, in other words, the answer to the *"why"* question.

- **Sensitivity analysis (SA)**: This kind of explanation can be done for both data features and training parameters. In the first case, the hypothetical output as a result of modifying (adding/removing/altering) the data features will be generated. In contrast, the second case deals with alterations in output as the result of modifications of the black-box's internal parameters (e.g., hidden layers in a deep neural network).
- **Direct what/how/why answers (DA)**: This category includes the most intuitive type of explanations, those which directly answer a question of type *what is, what happens if, why rather, how come* and other similar questions (see e.g., Holzinger, Malle, Saranti, & Pfeifer, 2021).

While being different in forms, we should emphasize that the first two structures mentioned above are special cases, or in other words, limited cases of the latter category. We have divided them into three groups to address the way they are being used in the XAI field. In the majority of works, often the origin of the feature importance and sensitivity analysis remains unmentioned that is, to which question (Why, How, etc.) are they responding, while in fact, many times explanations are rooted in a question which was raised by a specific user with a set of particular needs. We should emphasize that this is not the case for all types of explanations for instance those which are inspired by mathematical equations (See e.g. Layer-wise Relevance Propagation (LRP) Bach et al., 2015).

In our opinion, not elaborating on the choice of the explanation structure can damage the overall effectiveness of the XAI system, as the wrong structure will risk the knowledge transferring process as the final goal of any XAI system.

4.3. Explanation type

Working on explanations applied in information systems (IS), Hovorka, Germonprez, and Larsen (2008) proposes an expanded concept of explanations, arguing that the choice of explanation types depends on the reference disciplines through which research phenomena are understood and the research agenda is shaped. Below we briefly introduce these explanation types by providing a general form for each of them:

- **Covering-law(deductive-nomological) explanations**: "Whenever phenomenon X is observed to occur in the setting of conditions C, Y will be observed".
- **Statistical-relevance explanations**: "Based on empirical data, factors A, B and C contribute to the probability of Y by the amount of X".
- **Contrast-class explanation:** "In this context and given my purpose, why did X (rather than X*, X**, etc.) occur?"
- **Functional explanations**: "Identification of the mechanism by which desirable goal A ensures the continued existence of the phenomenon".

As the result of his survey, Miller, 2019 argues that a vast amount of such works (e.g., Dennett, 1989; Kass & Leake, 1987) are based on the four "modes of explanation" proposed by Aristotle, which are: Material (a substance which makes something), Formal (Form of something which its identity depends on), efficient(the proximal mechanism cause a change) and Final (the end goal of something). Miller further declares that *"explanations are contrastive"* and throughout his work, confronts it with *"complete explanations"* which, unlike the former, respond to straightforward **plain-fact** questions of the type *"why does object a have property P?"*, by listing the entire causal chain which results in the observed output.

Contrastive explanations (e.g., Lipton, 1990; Miller, 2019, 2021) are the natural response to a *why* questions, while some argue that *how* and *what* questions are also considered as such. Miller, 2019 points out that contrastive explanations provide a window in the questioner's mental model by showing their knowledge gap while at the same time, these explanations, with respect to the complete explanations discussed earlier, are more straightforward, more feasible and cognitively less demanding for both parties engaged in the explanation process. Different types of contrastive explanations are introduced in the literature and we go through some of them in the following part, but before doing so, we should mention what parts these different proposals of contrastive explanation have in common: Contrast class, fact and foil. For these concepts, we rely on the definition done by Robeer (2018): *"A contrast class F are all possible alternatives to a decision given the context (i.e., the range of values for a decision E). The fact is the actual decision $f \in F$, while the foil is any other member of the contrast class that is not fâ, i.e., $g \in F \setminus \{f\}$"*. As noted by McGill and Klein (1993), such definition of the counterfactuals, as the hypothetical outcome for event E, hold only for contrastive explanations while the same concept in the causality and its closely related concept, causation, is a "non-cause" in which the event-to-be-explained (Miller, 2019) does not occur. Van Bouwel and Weber (2002) goes further and introduces three types of contrastive questions:

- **P-contrast**: Why does object a have property P, rather than property Q?
- **O-contrast**: Why does object a have property P, while object b has property Q?
- **T-contrast**: Why does object a have property P at time t, but property Q at time t?

As Miller (2021) puts it, P-contrast – or the standard *"rather than"* question – happens within an object, O-contrast among objects themselves and T-contrast within an object over time. Furthermore, using the framework of Halpern and Pearl (2005), Miller categorizes the concept of P-contrast as "Alternative explanations" while labels O-contrast and T-contrast concepts as "Congruent questions", formalizing them in the Miller (2021).

Ylikoski (2007) provides another classification for contrastive questions as "incompatible" and "compatible" cases, while the former is when fact and foil are inconsistent and unlike the fact, the foil does not happen and is hypothetical (similar to P-contrast mentioned above), while in the latter case, fact and foil (or as he calls it, surrogate) are compatible and they both happen in diverse situations/times.

Finally, **Counterfactual** explanations answer to questions about the hypothetical outcome of a hypothetical event, or as Wachter, Mittelstadt, and Russell (2017) puts it, *"how the world would have to be different for a desirable outcome to occur"* (see e.g., Byrne, 2019; Verma, Dickerson, & Hines, 2020).

5. Message generation

Given the focus of our paper on the usage of natural language techniques in the XAI field, we provide a more detailed description of NLG techniques and dialogue systems as a presentation category and a sub-category of the text presentation.

The Explanation generator layer provides explanations of the black-box that cannot be delivered directly to the end-user as they are in a raw format, often using model-dependent notations. Hence, the role of this layer is to transform these raw outputs to explanations (e.g., *Messages*) that are comprehensible by the end-user.

5.1. Presentation technique

One of the less explored aspects of XAI is the presentation layer, where the explanations made by the explainer are transmitted to the end-user. The output of an XAI system can be multimodal, thus presenting natural language explanations and other content types. We grouped the presentation methods together: **Graphics/plots**, **Texts**, **Images** and **Reports**.

The choice of the representation methods depends on various interrelated factors. In our opinion, the most contributing ones include the ease of producing the representations (e.g., out-of-the-box solutions) and overlooking the importance of the presentation method on users' comprehension (Huysmans, Dejaeger, Mues, Vanthienen, & Baesens, 2011). In the following part, we briefly describe the typical presentation methods in the literature.

Graphics/Plots contain the most popular methods in the literature. Such popularity in our opinion is rooted in the presence of tools and the relative simplicity of generating such graphics. This group is mainly consists of the following types: *Bar plots*, *Line plots*, *Trees*, *heatmap plots*, *histograms*, *scatter plots* and *bubble plots*.

Bar Plot is the most used method in this category and can be further divided into Horizontal and Vertical Plots (Lou, Caruana, & Gehrke, 2012; Poulin et al., 2006; Ribeiro et al., 2016).

Line Plot vary from simple vertical bar plots to sophisticated custom plots made to represent a particular subject, often mixed with other types of methods like destiny plots or changing hue for adding additional attributes (Adler et al., 2018; Goldstein et al., 2015; Olden & Jackson, 2002).

Trees can be divided into two main categories: (i) Boolean Rules Trees which use the logic decision gates to classify records, and (ii) Decision Trees which utilize the Boolean decisions instead (Johansson, Niklasson, & König, 2004; Kato & Harada, 2014; Martens, Baesens, Van Gestel, & Vanthienen, 2007).

Heatmap Plot (not to be confused with Heatmap Images) are rather simple visual presentations that map a numeric value to its corresponding colour (Selvaraju et al., 2017; Zeiler & Fergus, 2014).

Histogram among the presented methods so far are the most technical methods as their interpretation might be complicated for the layman user without statistical knowledge (Baehrens et al., 2010; Olden & Jackson, 2002).

Scatter Plot , often boosted with other visualization methods, is used to map two or more dimensions into two or three-dimension space (Baehrens et al., 2010).

Bubble Plot, visually similar to scatter plot, can be considered as an augmented version of scatter plot and is often used to combine categorical and continuous values (Turner, 2016).

Images

Image-based presentations are considered more sophisticated with respect to the previous group (plots/graphics) and, at the same time, are more limited since they can be applied only if the target input is an image. The main types of this category are *image heatmap*, *saliency masking*, and *image manipulation*.

Image Heatmap, not to be confused with heat map plots, use an image as their base and add different layers of visualization, mostly coming from continuous data (Kim, Shah et al., 2015).

Saliency Masking is similar to image heatmaps as they utilize an image as their basis, but instead of adding heatmaps of values, they partially mask/cover the image to communicate a specific message (Bach et al., 2015; Ribeiro et al., 2016; Simonyan, Vedaldi, & Zisserman, 2014).

Image Manipulation, being the less sophisticated method in its family, image-manipulation consists of adding indicator shapes to an image in order to indicate a specific part of the image (Ribeiro, Singh, & Guestrin, 2018; Turner, 2016).

Reports

Although this family is close the *text* category, described below, reports have a more structured approach respect to texts and often are combined with other methods (e.g., graphics). The main techniques in this category are: *tabular reports*, *decision tables* and *graphical table reports*.

Tabular Report. The most basic method of this family, reports, conveys the desired message in a structured and direct manner (Henelius, Puolamäki, & Ukkonen, 2017; Poulin et al., 2006).

Decision Table. Like tabular reports, decision tables use the tabular structure, but since they solely represent the rules and mostly, no other info, they have less flexibility in the data types and other representations (Verbeke, Martens, Mues, & Baesens, 2011).

Graphical Table Report. This method, using tabular reports as the basis, integrates other methods in a very flexible way which allows one to customize the table based on the specific message desired to be communicated (Kim, Glassman, Johnson and Shah, 2015). **Texts**. This group contains methods that use the text as their basis. Notice that it does not mean that the output of these representations are necessarily expressed in natural language but indicates that the main message is conveyed through the text and not the other techniques mentioned before. The main textual representations are *rules*, *word annotations*, and *natural language* texts.

Another subcategory of AI which can mitigate the lack of explicit, symbolic representation of knowledge, i.e what prevents humans from fully comprehending black-boxes is Symbolic AI (Ciatto, Schumacher, Omicini, & Calvaresi, 2020). In this subcategory of AI the output can be in the textual/code (See Kitzelmann et al. (2006), Mao et al. (2018)). Inductive Logic Programming (ILP), a subfield of symbolic AI and more specifically a technique called Learning from Interpretation Transition (LFIT) can learn a propositional logic theory equivalent to a given black-box system under certain conditions (Ortega, Fierrez, Morales, Wang, & Ribeiro, 2021). Another example of Symbolic AI can be seen in the work of Malandri, Mercorio, Mezzanzanica, Nobani, and Seveso (2022b) that uses Binary Decision Diagram (BDD) to derive T-contrast explanations for text classifiers.

Natural Language Explanations. As part of text explanations, natural language explanations are text written in plain English or other human languages (see e.g., Hendricks et al., 2016; Hendricks, Hu, Darrell, & Akata, 2018a). Most of the approaches for generating explanations in natural language belong to the families of *NLG* and *Dialogue Systems*. Notice that, while the sentence generation task in dialogue systems is an application of NLG, they are more closely related to dialogue management since management and realization policies are usually learned together (Gatt & Krahmer, 2018). For this reason, we treat them as two different types of output.

NLG. In the seminal work of Reiter and Dale (1997), NLG is defined as *"The sub-field of artificial intelligence and computational linguistics that is concerned with the construction of computer systems that can produce meaningful texts in English or other human languages from some underlying non-linguistic representation of information"*. In essence, NLG is a branch of natural language processing (NLP) research (Cambria, Schuller, Xia, & White, 2016; Minaee et al., 2021; Zhao, Peng, Eger, Cambria, & Yang, 2019) focusing on the transformation of computer language to natural language. Traditionally, NLG tasks are divided into two main categories; data-to-text and text-to-text. As the name suggests, the data-to-text group deals with generating natural language mainly from numerical data. As examples of this category, we can mention Robo-Journalism or automatic reporting (e.g., automatic weather forecast Sripada, Reiter, & Davy, 2003 and sports event reports Chen & Mooney, 2008). Text-to-text category, on the other hand, covers a wider and somehow more significant applications like machines translation (e.g., Devlin et al., 2014; Koehn et al., 2007), text summarization and simplification (Webber, Egg, & Kordoni, 2012) and paraphrasing (Androutsopoulos & Malakasiotis, 2010). While the mentioned categories of NLG are the major players in the NLG field, in the past decade, another group, vision-to-text, has emerged, mainly thanks to the proliferation of the Deep Neural Network methods. Although this category is not yet as mature as the methods mentioned above, there already exist numerous applications like image captioning (Yang, Tang, Zhang, & Cai, 2019) and the generation of natural language explanations using Deep Learning techniques (e.g. Chang, Harper, and Terveen (2016), Costa, Ouyang, Dolog, and Lawlor (2018) and Ehsan et al. (2019)) As Mariotti et al. (2020) clarifies, NLG works can also be divided considering the technology used for generations of the text: *Template-based*, which structure templates that present the output in textual form and *End-to-end* generation which utilizes large humanly labelled data-to-text corpora.

Rules. Rules are simply a list of decision rules written in natural language (Letham, Rudin, McCormick, Madigan, et al., 2015; Martens et al., 2007; Zhou, Jiang, & Chen, 2003).

Word annotation. Word annotation is the fastest and simplest method of the text category, in which a message is conveyed by highlighting or changing the colour of a specific part of the text (Kim, Glassman et al., 2015; Lei, Barzilay, & Jaakkola, 2016; Liu, Zhang, & Gulla, 2020; Ribeiro et al., 2016).

Dialogue Systems. Essentially, a dialogue system is a system that enables the conversation between two parties. Neither the term *dialogue system* nor its definition, however, have a clear consensus among the researchers. While there are various alternatives for this term, we can mention *conversational Agent*, *Conversational User Interface* and *Chatbots* are the most commonly used in academia and business (Li, Shao, Ji, & Cambria, 2022; Ma, Nguyen, Xing, & Cambria, 2020; Xu, Peng, Xie, Cambria, Zhou, & Zheng, 2020; Young et al., 2018; Young, Pandelea, Poria, & Cambria, 2020). Although such systems have been around for the past fifty years, their usage as a presentation layer in XAI systems is minimal. Hilton (1990) defines an explanation as *"someone explains something to someone else"*, which emphasizes the conversation form of a causal explanation. What we saw in the proposed roadmap until this point was generating explanations using context and presenting them as a fixed explanation. An alternative to such presentations is to communicate the message (explanation) as a part of a conversation between the system (explainer) and the user (explainee).

6. Keeping the roadmap up-to-date

One of the significant limitations of survey studies is that they rapidly become obsolete as soon as the research on the topic advances. Aiming at overcoming this issue, Qian et al. (2021), propose an interactive browser-based system called XNLP.[2] which *synthesizes the state of the field at different levels of abstractions and from different perspectives* Although this tool, in our opinion, brings

[2] https://xainlp2020.github.io/xainlp/home.

Table 1
Mapping selected papers to our roadmap. *(Example and Benchmark)* → Not provided/used: □, Provided/used once: ⧉, Provided/used multiple times: ■ ; *(Dataset)* → Not mentioned: 👁̸, Private dataset: 🔒, Public dataset: 🔓; *(Code)* → Not provided: 🔑̸, Provided no documentation: git, Provided with documentation: git; *(Rest of features)* → Not mentioned: ○, Mentioned but not applied: ◉, Applied: ● (see below-mentioned reference for further information: Amarasinghe & Manic, 2019; Donadello & Dragoni, 2021; Hendricks, Hu, Darrell, & Akata, 2018b; Sreedharan, Srivastava, & Kambhampati, 2021)

	Experiments				Evaluation		Explanation goal			Audience			Explanator type				Structure			Explanation type				Presentation				
Paper	Example	Benchmark	Dataset	Code	User-Evaluation	Metric	Model Explanation	Outcome Explanation	Model Inspection	end user	Developer	Decision-Maker	Decision Tree	Decision Rule	Salient Masks	Feature Inspection	Feature Importance	Sensitivity Analysis	Direct Answers	Plain-fact	Contrastive	Counterfactual	Graphic	Image	Rule	Text	Dialogue System
Costa et al. (2018)	■	□	🔓	🔑̸	○	●	○	○	●	●	○	○	○	○	○	◉	○	○	●	●	○	○	○	○	○	●	○
Ehsan et al. (2019)	■	□	👁̸	🔑̸	●	●	○	●	○	●	○	○	○	○	○	○	○	○	●	●	○	○	●	○	○	●	○
Chang et al. (2016)	■	⧉	🔓	🔑̸	●	○	○	○	●	●	○	○	○	○	○	○	●	○	●	●	○	○	○	○	○	●	○
Hendricks et al. (2018a)	■	□	🔓	🔑̸	○	●	○	●	○	○	○	○	○	○	●	○	○	○	●	○	○	●	○	●	○	●	○
Alonso and Bugarín (2019)	■	□	🔓	git	◉	○	●	●	○	●	●	◉	●	○	○	○	○	○	○	●	○	○	●	○	○	●	○
Sokol and Flach (2018)	⧉	□	🔓	🔑̸	○	○	○	●	○	○	○	○	●	○	○	○	○	◉	●	●	●	●	○	○	○	○	●
Core et al. (2006)	■	□	👁̸	🔑̸	○	○	○	●	○	●	○	○	○	◉	○	○	○	○	●	●	○	○	○	○	○	○	●
Rosenthal, Selvaraj, and Veloso (2016)	■	□	👁̸	🔑̸	○	●	○	●	○	●	●	○	○	○	○	○	○	○	●	●	○	○	○	○	○	●	○
Amarasinghe and Manic (2019)	■	□	🔓	🔑̸	○	●	●	○	○	●	○	○	○	◉	○	○	○	◉	○	●	○	○	○	○	○	●	○
Hohman et al. (2019)	■	□	👁̸	🔑̸	○	○	●	●	●	●	●	●	○	○	○	○	●	○	○	●	○	●	●	○	○	●	○
Park et al. (2018)	■	⧉	🔓	🔑̸	●	●	○	●	○	●	○	○	○	○	●	○	○	○	●	●	○	○	○	●	○	●	○
Malandri, Mercorio, Mezzanzanica, Nobani, and Seveso (2022a)	■	⧉	🔓	🔑̸	○	●	●	○	○	●	○	◉	●	●	○	○	○	○	●	○	●	○	●	○	●	●	○
Zhao, Huang, Huang, Robu, and Flynn (2021)	■	⧉	🔓	git	○	●	○	●	○	●	○	○	○	○	●	●	●	○	○	●	○	○	●	●	○	○	○
Donadello and Dragoni (2021)	■	□	🔓	git	○	○	○	●	○	●	○	○	●	○	○	●	●	○	●	●	○	○	○	○	○	●	○
Hendricks et al. (2016)	■	⧉	🔓	🔑̸	●	●	○	●	○	●	○	○	○	○	○	●	●	○	○	●	○	○	○	○	○	●	○
Hendricks et al. (2018b)	■	⧉	🔓	🔑̸	●	●	○	●	○	●	○	○	○	○	○	●	●	○	○	●	○	●	○	●	○	●	○
Sreedharan et al. (2021)	■	⧉	🔓	🔑̸	●	●	○	●	○	●	○	○	○	○	○	●	○	○	●	○	●	○	○	○	○	●	○

value to the XAI community by providing a dynamic hub of recent works, it lacks a feature that directs researchers to the most related works to their field of research.

To bridle this limitation, we propose to model the roadmap depicted in Fig. 1 as a multi-criteria-decision-making (MCDM) problem where columns of Table 1 are criteria whilst the rows are the alternatives on which decide. Hence, the decision goal is *to decide which is the most suitable XAI system that makes use of natural language explanations*. Modelling such a decision as an MCDM problem allows deciding, taking into account the user needs (criteria) and their relative importance of them (weight of criteria).

6.1. Multi-criteria decision-making at a glance

In essence, MCDM refers to a set of methods that allows constructing a global preference relation for a set of alternatives to be evaluated by using several criteria. A literature review on MCDM falls out of the scope of this paper; the reader can refer to Figueira, Greco, and Ehrgott (2005) for a survey. The MDCM approaches are able to deal with dependence amongst criteria (e.g., ANP Saaty,

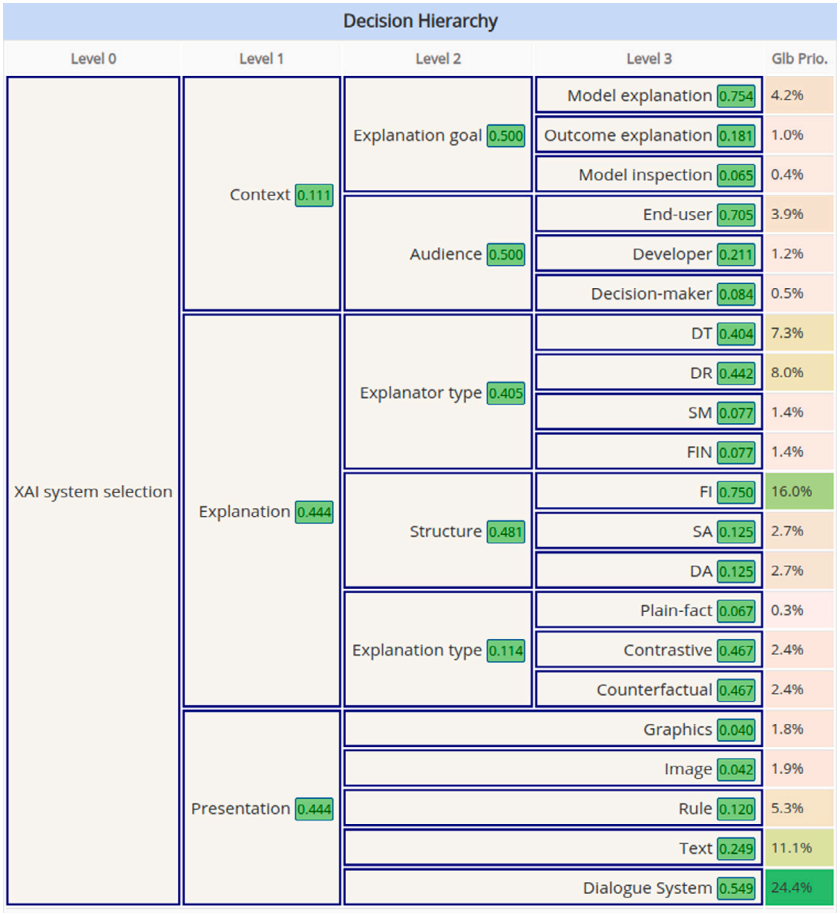

Fig. 2. The AHP hierarchy built from Table 1 weighted by a user. Any user can contribute weighting the hierarchy at https://tinyurl.com/XAI-NLG-AHP.

2004), conflicting criteria (ELECTRE), synthesize compromise solutions (TOPSIS), as well as to deal with uncertainty over the judgments (Fuzzy sets theory applied to the previous methods). In our work, we use the *Analytic Hierarchy Process* (AHP) (Saaty, 1987), as it is beneficial for evaluating complex multi-attribute alternatives involving subjective criteria to capture stakeholders' knowledge of phenomena under study. AHP consists of the following main steps.

(i) Build up the criteria/alternatives tree. In this step, the criteria that compose the decision problem are identified and organized hierarchically so that a criterion may have sub-criteria, and so on. The leaves of this tree are the alternatives that the decision process aims at selecting. Our hierarchy of criteria is drawn following Table 1.

(ii) Pairwise Comparison of Criteria. In this step, the users are required to perform a pairwise comparison of each criterion at each level of the hierarchy, and the results are collected in a matrix summarizing the local priorities for each domain expert. The main intuition here is that it is easier (and more accurate) to compare the importance of two criteria at a time than simultaneously evaluating all of them. There are two characteristics of AHP that deserve to be highlighted. First, the same preference scale, i.e., the Saaty's Scale (Saaty, 2004), is used to evaluate both (quantitative and qualitative) criteria and alternatives. Second, the expert does not provide any absolute numerical judgment but a comparative evaluation, which is more familiar to people. Comparisons are recorded in a positive reciprocal matrix, in which a_{ij} represents the comparison between element i and j.

The rationale of the relationship $a_{ji} = 1/a_{ij}$ is that if A is four times more important than B, then B is 1/4 important with respect to A. Thus, if the matrix is perfectly consistent, the transitivity rule is satisfied for all the comparisons, namely $a_{ij} = a_{ik} \cdot a_{kj}$. Intuitively, it is expected that if A is *moderate important* (3) than B, and *B* is *weak important* (2) than C, thus a consistent judgment would have that A is $3 \cdot 2 = 6$ *strong important* than C. As inconsistencies are natural in human judgments, AHP provides the consistency ratio to the final user. It was proved that inconsistencies in answers could be tolerated if the consistency ratio remains within a small interval, that is 10% (Saaty, 2004).

At the end of this process, a weighted hierarchy that encodes the user preference is obtained, as in Fig. 2. Notice AHP allows group decision-making by averaging judgments into one unified weighted hierarchy.

(iii) Synthesize Global Priorities of Alternatives. The last step requires synthesizing the global priorities (i.e., the priority vector) from the pairwise comparisons to determine the ranking of *alternatives*, taking into account the user judgments computed in

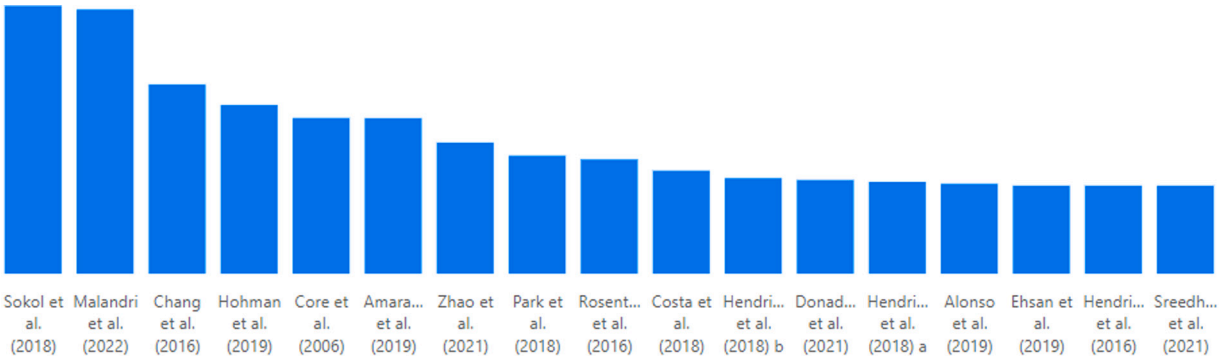

Fig. 3. The paper ranking based on the hierarchy shown in Fig. 1.

the previous step. Mathematically speaking, the priority vector is the solution of an Eigenvalue problem over the matrix previously introduced. The results of the pairwise comparisons are arranged in a matrix. The matrix's first (dominant) normalized right Eigenvector gives the ratio scale (weighting), while the Eigenvalue determines the consistency ratio. At the end of this step, a list of alternatives ranked is provided, as in Fig. 3; the figure shows the paper rankings user got based on the created hierarchy, after a pairwise comparison of papers. In our approach, once the weighted hierarchy of criteria is obtained, the pairwise evaluation of *alternatives* is automatically performed drawing from Table 1, by normalizing the values on the Saaty's scale. One might note that the user assigned 44% of importance to both the Presentation and Explanation layers, whilst the Context account for the 11%. Looking at global priorities, having a *Dialogue System* accounts for the 24.4% on the final decision, as well as being able to provide Feature Importance (FI) accounts for the 16% globally, more than having *text* (11.1%).

Based on these results, the paper that better fits the preferences of the decision hierarchy in Fig. 3 is Sokol and Flach (2018) with the consolidated weight of 14.3%. This paper can explain this output is one of two using *dialogue systems*, which have the highest priority in the hierarchy defined by the user. The reason why the output is not (Core et al., 2006) – the other paper using a dialogue system – is that, unlike the latter, Sokol and Flach (2018) uses Decision Trees which is another relevant criterion in the hierarchy. In essence, AHP allows one to capture and keep track of the reason behind the decision, taking into account the relative importance of the XAI characteristics.

One should note that users are able to update the roadmap criteria and add new alternatives (papers) to adapt the framework and update it based on their specific needs. In order to demonstrate how the framework transforms specific users' needs into paper rankings, in the following part we provide three simulated cases including the initial need, hierarchy weights of criteria and paper rankings.

Case 1 Working with a dataset which has both images and separate features the researcher's goal is to classify EMG hand movement. To do so, the researcher wants to be able to explain individual outcomes of the black-box to final users. Such explanations should be able to satisfy both plain-fact (why questions) and contrastive (why not or why this and not that) questions of end-users, using either rules or images. It is also acceptable to receive such explanations through a conversation.

Case 2 The researcher's goal is to classify job vacancies and have a clear understanding of the characteristics of the classification. The explanations are destinated for the decision-makers who act upon the obtained results. The preferred mode for explanations to be conveyed are decision trees and rules while the preferred presentations are through rules, graphics or natural language.

Case 3 Having tabular data of credit landing the researcher aims to inspect the model and assess its fairness. The target of the explanations are the developers and the most efficient way for them to comprehend and act upon them is through decision rules and trees and in form of plain-fact and counterfactual explanations. Similarly, the preferred modality of presentation for this specific target is textual or rule explanations.

Discussion Figs. 4–6 show the decision hierarchy and the resulting ranking of papers based on these preferences. As it can be observed, the paper ranking accurately reflects the preferences of researchers in terms of explanation goal, audiences and presentation choices. For instance, in case 1, the researcher working with image data prefers to receive the explanations in a plain-fact manner and thorough images, which is aligned with the paper ranking provided by the tool, with Zhao et al. (2021) as the first paper. Similarly, in case 2, where the preferred explanations are rules and natural language, the top-ranked papers are Hohman et al. (2019) and Malandri et al. (2022a) which provide such explanations. Finally, in case 3 the users is a developer that need a XAI algorithm able to process tabular data and to provide explanations in the form of rules and trees, like the proposed ones. Note that, given that both case 2 and case 3 require decision rules and trees as explanations, the first three suggestions point to the same methods, while the following differs. For instance, the verbalization model proposed by Rosenthal et al. (2016) is not designed for text classification. Therefore, it appears as fourth best-fit for case 3, but it is much lower in the ranking for case 2. Those examples show the usefulness of the proposed approach, which helps the users in narrowing the search among the vast range of available XAI methods, pointing the one that better fits their need and tasks.

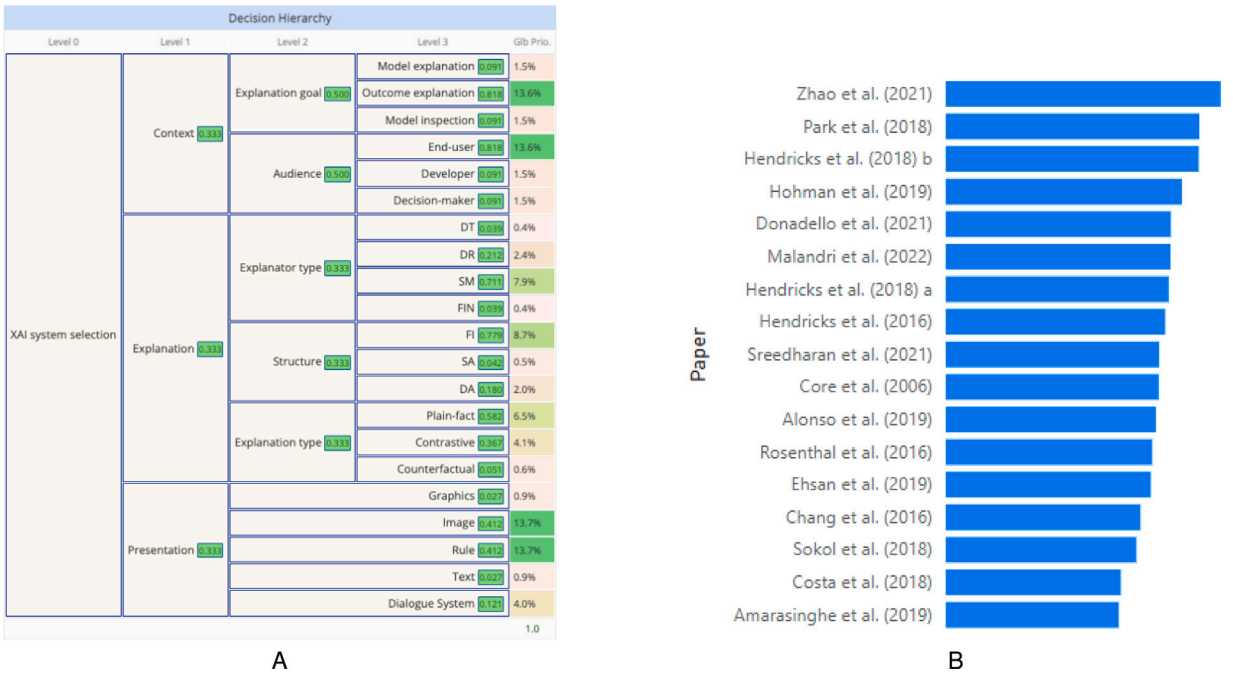

Fig. 4. Hierarchy (A) and paper ranking (B) of case 1.

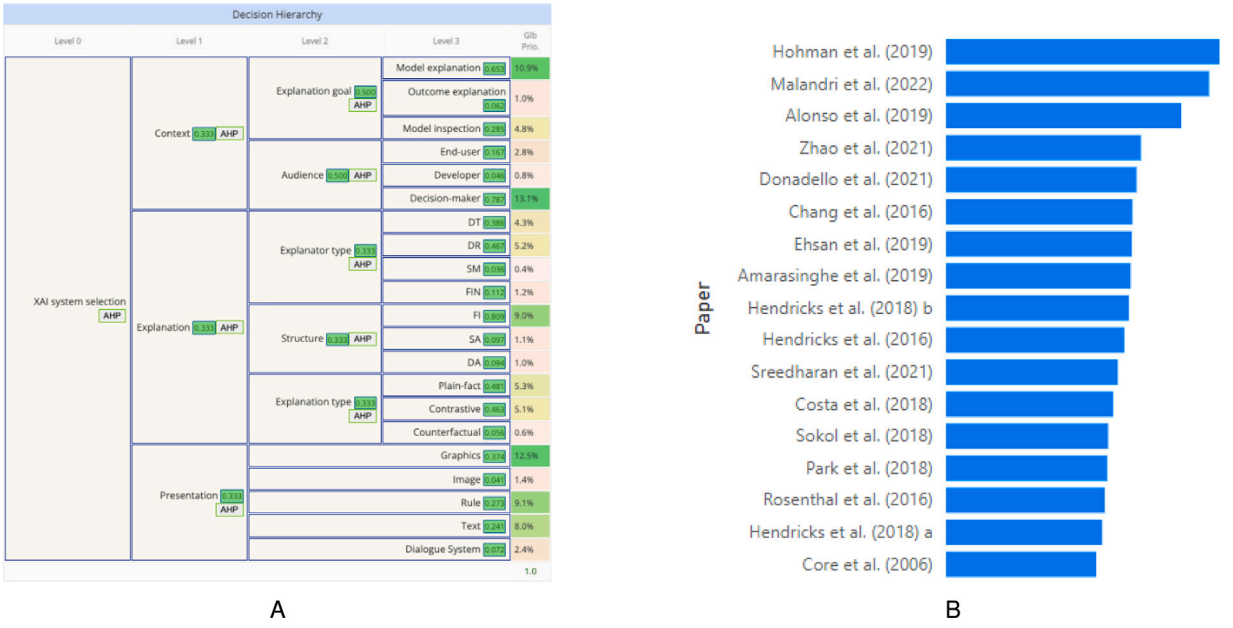

Fig. 5. Hierarchy (A) and paper ranking (B) of case 2.

7 Conclusion and future work

In this work, we surveyed XAI methods that make use of natural language to provide explanations (i.e., either through *text* or *dialogue systems*) of what has been done, what is done right now, what will be done next and unveil the information the actions are based on. We considered 70 XAI papers that encompass natural language explanations published in top-tier conferences and journals between 2006 and 2021. We proposed a roadmap to analyse the whole explanation process, from the black-box model to the final user, going through three main layers: context definition, explanation generation and message generation. Furthermore, we mapped the main methods in XAI literature to a list of the characteristic of each layer. As a further contribution, we modelled the

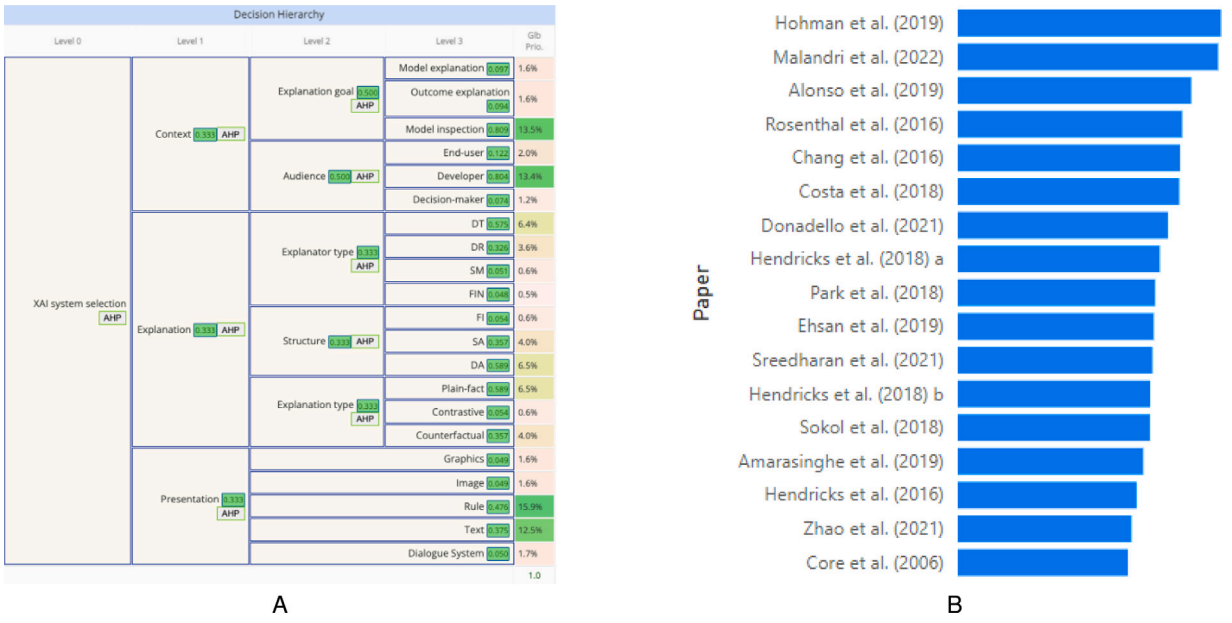

Fig. 6. Hierarchy (A) and paper ranking (B) of case 3.

roadmap as a multi-criteria-decision-making problem and employed the analytic hierarchy process to select the XAI technique that better fits user preferences. Finally, to foster knowledge-sharing among XAI researchers, the model allows for adding new papers on top of the ones already reviewed here to keep the evaluation framework updated over time. As for future works, there is a need for an in-depth study of explanations generated within Human–Machine Teams (HMT) (see e.g. Paleja, Ghuy, Ranawaka Arachchige, Jensen, and Gombolay (2021)) which go beyond interactive dialogues, where the explainer is not considered a mere tool but a team member which is present in every step of the decision-making process and is capable to predict and act upon alterations in the context and environment parameters.

CRediT authorship contribution statement

Erik Cambria: Supervision, Writing – review & editing. **Lorenzo Malandri:** Formal analysis, Writing – original draft, Writing – review & editing, Validation. **Fabio Mercorio:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision. **Mario Mezzanzanica:** Supervision, Writing – original draft, Project administration. **Navid Nobani:** Conceptualization, Software, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization.

Data availability

The following link provides a GitHub repository with instructions for using the AHP tool: https://bit.ly/3S6zTp8. Please consider that the repository will be enriched periodically.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *CHI*.

Adler, P., Falk, C., Friedler, S. A., Nix, T., Rybeck, G., Scheidegger, C., et al. (2018). Auditing black-box models for indirect influence. *Knowledge and Information Systems*, *54*(1), 95–122.

Alonso, J. M., & Bugarín, A. (2019). ExpliClas: Automatic generation of explanations in natural language for weka classifiers. In *FUZZ-IEEE*.

Alonso, J. M., Ramos-Soto, A., Reiter, E., & van Deemter, K. (2017). An exploratory study on the benefits of using natural language for explaining fuzzy rule-based systems. In *FUZZ-IEEE*.

Amarasinghe, K., & Manic, M. (2019). Explaining what a neural network has learned: Toward transparent classification. In *FUZZ-IEEE*.

Androutsopoulos, I., & Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, *38*.

Apicella, A., Isgrò, F., Prevete, R., & Tamburrini, G. (2019). Contrastive explanations to classification systems using sparse dictionaries. In *ICIAP*. Springer.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One*, *10*(7), Article e0130140.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & Müller, K.-R. (2010). How to explain individual classification decisions. *Journal of Machine Learning Research*.

Bennetot, A., Donadello, I., Qadi, A. E., Dragoni, M., Frossard, T., Wagner, B., et al. (2021). A practical tutorial on explainable AI techniques. arXiv preprint arXiv:2111.14260.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., et al. (2020). Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 648–657).

Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*.

Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*.

Byrne, R. M. (2019). Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In *IJCAI* (pp. 6276–6282).

Cambria, E., Hussain, A., Havasi, C., & Eckl, C. (2009). Common sense computing: From the society of mind to digital intuition and beyond. In J. Fierrez, J. Ortega, A. Esposito, A. Drygajlo, & M. Faundez-Zanuy (Eds.), *Lecture Notes in Computer Science*: *Vol. 5707*, *Biometric ID Management and Multimodal Communication* (pp. 252–259). Berlin Heidelberg: Springer.

Cambria, E., Liu, Q., Decherchi, S., Xing, F., & Kwok, K. (2022). SenticNet 7: A commonsense-based neurosymbolic ai framework for explainable sentiment analysis. In *LREC*.

Cambria, E., Schuller, B., Xia, Y., & White, B. (2016). New avenues in knowledge bases for natural language processing. *Knowledge-Based Systems*, *108*, 1–4.

Caruana, R., Lundberg, S., Ribeiro, M. T., Nori, H., & Jenkins, S. (2020). Intelligible and explainable machine learning: Best practices and practical challenges. In *ACM-SIGKDD* (pp. 3511–3512).

Chang, S., Harper, F. M., & Terveen, L. G. (2016). Crowd-based personalized natural language explanations for recommendations. In *RecSys*.

Chaves, A. P., & Gerosa, M. A. (2020). How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction*, 1–30.

Chen, D. L., & Mooney, R. J. (2008). Learning to sportscast: a test of grounded language acquisition. In *ICML*.

Ciatto, G., Schumacher, M. I., Omicini, A., & Calvaresi, D. (2020). Agent-based explanations in AI: towards an abstract framework. In *International workshop on explainable, transparent autonomous agents and multi-agent systems* (pp. 3–20). Springer.

Core, M. G., Lane, H. C., Van Lent, M., Gomboc, D., Solomon, S., & Rosenberg, M. (2006). Building explainable artificial intelligence systems. In *AAAI*.

Costa, F., Ouyang, S., Dolog, P., & Lawlor, A. (2018). Automatic generation of natural language explanations. In *IUI*.

De Gennaro, M., Krumhuber, E. G., & Lucas, G. (2020). Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood. *Frontiers in Psychology*, *10*, 3061.

Dennett, D. C. (1989). *The intentional stance*. MIT Press.

Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R., & Makhoul, J. (2014). Fast and robust neural network joint models for statistical machine translation. In *ACL*.

Donadello, I., & Dragoni, M. (2021). Bridging signals to natural language explanations with explanation graphs.

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. In T. R. Besold, & O. Kutz (Eds.), *AI*IA*.

Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., & Riedl, M. O. (2019). Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *IUI*.

Figueira, J., Greco, S., & Ehrgott, M. (2005). Multiple criteria decision analysis: state of the art surveys.

Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, *61*, 65–170.

Giabelli, A., Malandri, L., Mercorio, F., Mezzanzanica, M., & Seveso, A. (2020). NEO: A tool for taxonomy enrichment with new emerging occupations. In *ISWC* (pp. 568–584).

Gkatzia, D., Lemon, O., & Rieser, V. (2016). Natural language generation enhances human decision-making with uncertain information. In *ACL*.

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*.

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018). Local rule-based explanations of black box decision systems. arXiv preprint arXiv:1805.10820.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5).

Hall, M., Harborne, D., Tomsett, R., Galetic, V., Quintana-Amate, S., Nottle, A., et al. (2019). A systematic method to understand requirements for explainable AI (XAI) systems. In *IJCAI*.

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science*, *56*(4).

Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., & Darrell, T. (2016). Generating visual explanations. In *ECCV*. Springer.

Hendricks, L. A., Hu, R., Darrell, T., & Akata, Z. (2018a). Generating counterfactual explanations with natural language. In *ICML WHI*.

Hendricks, L. A., Hu, R., Darrell, T., & Akata, Z. (2018b). Grounding visual explanations. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 264–279).

Henelius, A., Puolamäki, K., & Ukkonen, A. (2017). Interpreting classifiers through attribute interactions in datasets. In *ICML WHI*.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, *107*(1).

Hoffman, R. R., Klein, G., & Mueller, S. T. (2018). Explaining explanation for "explainable AI". In *Proceedings of the human factors and ergonomics society annual meeting, Vol. 62* (pp. 197–201). Los Angeles, CA: SAGE Publications Sage CA.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608.

Hohman, F., Srinivasan, A., & Drucker, S. M. (2019). TeleGam: Combining visualization and verbalization for interpretable machine learning. In *VIS*. IEEE.

Holzinger, A., Carrington, A., & Müller, H. (2020). Measuring the quality of explanations: the system causability scale (SCS). *KI-Künstliche Intel.*, *34*(2), 193–198.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(4), Article e1312.

Holzinger, A., Malle, B., Saranti, A., & Pfeifer, B. (2021). Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI. *Information Fusion*, *71*, 28–37.

Hovorka, D. S., Germonprez, M., & Larsen, K. R. (2008). Explanation in information systems. *International Surgery Journal*.

Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., & Baesens, B. (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, *51*(1), 141–154.

Jentzsch, S. F., Höhn, S., & Hochgeschwender, N. (2019). Conversational interfaces for explainable AI: a human-centred approach. In *International workshop on explainable, transparent autonomous agents and multi-agent systems* (pp. 77–92). Springer.

Johansson, U., Niklasson, L., & König, R. (2004). Accuracy vs. comprehensibility in data mining models. In *Proceedings of the seventh international conference on information fusion, Vol. 1* (pp. 295–300). Citeseer.

Johs, A. J., Agosto, D. E., & Weber, R. O. (2020). Qualitative investigation in explainable artificial intelligence: A bit more insight from social science. arXiv preprint arXiv:2011.07130.

Kass, A., & Leake, D. (1987). *Types of explanations*: *Technical report*, Yale Univ New Haven Ct Dept of Computer Science.

Kato, H., & Harada, T. (2014). Image reconstruction from bag-of-visual-words. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 955–962).

Kenny, E. M., Ford, C., Quinn, M., & Keane, M. T. (2021). Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence*, *294*, Article 103459.

Kim, B., Glassman, E., Johnson, B., & Shah, J. (2015). iBCM: Interactive Bayesian case model empowering humans via intuitive interaction.
Kim, B., Shah, J., & Doshi-Velez, F. (2015). Mind the gap: a generative approach to interpretable feature selection and extraction. In *NIPS*.
Kitzelmann, E., Schmid, U., Olsson, R., & Kaelbling, L. P. (2006). Inductive synthesis of functional programs: An explanation based generalization approach. *Journal of Machine Learning Research*, *7*(2).
Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *ACL*.
Krahmer, E., & Theune, M. (2010). *Empirical methods in natural language generation: data-oriented methods and empirical evaluation, Vol. 5790*. Springer.
Kulesza, T., Stumpf, S., Burnett, M., & Kwan, I. (2012). Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *SIGCHI*.
Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., et al. (2021). What do we want from explainable artificial intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, *296*, Article 103473.
Lecue, F. (2020). On the role of knowledge graphs in explainable AI. *Semantic Web*, *11*(1), 41–51.
Lei, T., Barzilay, R., & Jaakkola, T. (2016). Rationalizing neural predictions. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 107–117).
Letham, B., Rudin, C., McCormick, T. H., Madigan, D., et al. (2015). Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics*, *9*(3), 1350–1371.
Li, Y., Pan, Q., Wang, S., Yang, T., & Cambria, E. (2018). A generative model for category text generation. *Information Sciences*, *450*, 301–315.
Li, W., Shao, W., Ji, S., & Cambria, E. (2022). BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis. *Neurocomputing*, *467*, 73–82.
Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplements*, *27*, 247–266.
Liu, P., Zhang, L., & Gulla, J. A. (2020). Dynamic attention-based explainable recommendation with textual and visual fusion. *Information Processing & Management*, *57*(6), Article 102099.
Lou, Y., Caruana, R., & Gehrke, J. (2012). Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 150–158).
Lucic, A., Haned, H., & de Rijke, M. (2020). Why does my model fail? Contrastive local explanations for retail forecasting. In *ACM FAccT*.
Ma, Y., Nguyen, K. L., Xing, F., & Cambria, E. (2020). A survey on empathetic dialogue systems. *Information Fusion*, *64*, 50–70.
Malandri, L., Mercorio, F., Mezzanzanica, M., Nobani, N., & Seveso, A. (2022a). ContrXt: Generating contrastive explanations from any text classifier. *Information Fusion*, *81*, 103–115.
Malandri, L., Mercorio, F., Mezzanzanica, M., Nobani, N., & Seveso, A. (2022b). The good, the bad, and the explainer: a tool for contrastive explanations of text classifiers. In *IJCAI*.
Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., & Wu, J. (2018). The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *International conference on learning representations*.
Mariotti, E., Alonso, J. M., & Gatt, A. (2020). Towards harnessing natural language generation to explain black-box models. In *NL4XAI*.
Martens, D., Baesens, B., Van Gestel, T., & Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research*, *183*(3), 1466–1476.
McClure, J. (2002). Goal-based explanations of actions and outcomes. *European Review of Social Psychology*, *12*(1), 201–235.
McGill, A. L., & Klein, J. G. (1993). Contrastive and counterfactual reasoning in causal judgment. *Journal of Personality and Social Psychology*, *64*(6).
Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*.
Miller, T. (2021). Contrastive explanation: A structural-model approach. *The Knowledge Engineering Review*, *36*.
Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning based text classification: A comprehensive review. *ACM Computing Surveys*, *54*(3), Article 62.
Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. In *ACM FAccT*.
Muller, H., Mayrhofer, M. T., Van Veen, E.-B., & Holzinger, A. (2021). The ten commandments of ethical medical AI. *Computer*, *54*(07), 119–123.
Olden, J. D., & Jackson, D. A. (2002). Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling*, *154*(1–2), 135–150.
Ortega, A., Fierrez, J., Morales, A., Wang, Z., & Ribeiro, T. (2021). Symbolic AI for XAI: Evaluating LFIT inductive programming for fair and explainable automatic recruitment. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 78–87).
Paleja, R., Ghuy, M., Ranawaka Arachchige, N., Jensen, R., & Gombolay, M. (2021). The utility of explainable ai in ad hoc human-machine teaming. *Advances in Neural Information Processing Systems*, *34*, 610–623.
Park, D. H., Hendricks, L. A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., et al. (2018). Multimodal explanations: Justifying decisions and pointing to the evidence. In *CVPR*.
Payrovnaziri, S. N., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J. H., et al. (2020). Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *Journal of the American Medical Informatics Association*, *27*(7), 1173–1185.
Poulin, B., Eisner, R., Szafron, D., Lu, P., Greiner, R., Wishart, D. S., et al. (2006). Visual explanation of evidence with additive classifiers. In *Proceedings of the national conference on artificial intelligence, Vol. 21* (p. 1822). Menlo Park, CA; Cambridge, MA; London: AAAI Press; MIT Press, 1999.
Przybyła, P., & Soto, A. J. (2021). When classification accuracy is not enough: Explaining news credibility assessment. *Information Processing & Management*, *58*, Article 102653.
Qian, K., Danilevsky, M., Katsis, Y., Kawas, B., Oduor, E., Popa, L., et al. (2021). XNLP: A living survey for XAI research in natural language processing. In *26th international conference on intelligent user interfaces* (pp. 78–80).
Raman, V., Lignos, C., Finucane, C., Lee, K. C., Marcus, M. P., & Kress-Gazit, H. (2013). Sorry dave, i'm afraid I can't do that: Explaining unachievable robot tasks using natural language.. In *Robotics: science and systems*.
Reiter, E., & Dale, R. (1997). Building applied natural language generation. *Natural Language Engineering*, *3*(1).
Reiter, E., & Dale, R. (2000). *Studies in natural language processing, Building natural language generation systems*.
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. In *ACM SIGKDD*.
Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 32*.
Ribera, M., & Lapedriza, A. (2019). Can we do better explanations? A proposal of user-centered explainable AI. In *IUI workshops*.
Robeer, M. J. (2018). *Contrastive explanation for machine learning* (Master's thesis).
Rosenthal, S., Selvaraj, S. P., & Veloso, M. M. (2016). Verbalization: Narration of autonomous robot experience. In *IJCAI*.
Saaty, R. W. (1987). The analytic hierarchy process—what it is and how it is used. *Mathematical Modelling*, *9*(3–5), 161–176.
Saaty, T. L. (2004). Fundamentals of the analytic network process—multiple networks with benefits, costs, opportunities and risks. *Journal of Sport & Social Issues*.
Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618–626).
Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In workshop at international conference on learning representations*. Citeseer.
Sokol, K., & Flach, P. A. (2018). Conversational explanations of machine learning predictions through class-contrastive counterfactual statements. In *IJCAI*.

Sokol, K., & Flach, P. (2020). LIMEtree: Interactively customisable explanations based on local surrogate multi-output regression trees. arXiv preprint arXiv:2005.01427.

Sreedharan, S., Srivastava, S., & Kambhampati, S. (2021). Using state abstractions to compute personalized contrastive explanations for ai agent behavior. *Artificial Intelligence*, *301*, Article 103570.

Sripada, S., Reiter, E., & Davy, I. (2003). SumTime-Mousam: Configurable marine weather forecast generator. *Expert Update*, *6*(3).

Stöger, K., Schneeberger, D., & Holzinger, A. (2021). Medical artificial intelligence: the European legal perspective. *Communications of the ACM*, *64*(11), 34–36.

Sturm, I., Lapuschkin, S., Samek, W., & Müller, K.-R. (2016). Interpretable deep neural networks for single-trial EEG classification. *Journal of Neuroscience Methods*, *274*, 141–145.

Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. In *ICML*.

Turner, R. (2016). A model explanation system. In *2016 IEEE 26th international workshop on machine learning for signal processing (MLSP)* (pp. 1–6). IEEE.

Van Bouwel, J., & Weber, E. (2002). Remote causes, bad explanations? *Journal for the Theory of Social Behaviour*, *32*(4).

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*, *38*(3), 2354–2364.

Verma, S., Dickerson, J., & Hines, K. (2020). Counterfactual explanations for machine learning: A review. arXiv preprint arXiv:2010.10596.

Vilone, G., & Longo, L. (2020). Explainable artificial intelligence: a systematic review. arXiv preprint arXiv:2006.00093.

Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, *76*, 89–106.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *The Harvard Journal of Law & Technology*, *31*, 841.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable AI. In *CHI*.

Webber, B., Egg, M., & Kordoni, V. (2012). Discourse structure and language technology. *Natural Language Engineering*, *18*(4), 437.

Werner, C. (2020). Explainable AI through rule-based interactive conversation. In *EDBT/ICDT*.

Xu, H., Peng, H., Xie, H., Cambria, E., Zhou, L., & Zheng, W. (2020). End-to-end latent-variable task-oriented dialogue system with exact log-likelihood optimization. *World Wide Web*, *23*, 1989–2002.

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). Explainable AI: A brief survey on history, research areas, approaches and challenges. In *CCF international conference on natural language processing and Chinese computing* (pp. 563–574). Springer.

Yang, X., Tang, K., Zhang, H., & Cai, J. (2019). Auto-encoding scene graphs for image captioning. In *CVF*.

Ylikoski, P. (2007). The idea of contrastive explanandum. In *Rethinking explanation*. Springer.

Young, T., Cambria, E., Chaturvedi, I., Zhou, H., Biswas, S., & Huang, M. (2018). Augmenting end-to-end dialogue systems with commonsense knowledge. In *AAAI* (pp. 4970–4977).

Young, T., Pandelea, V., Poria, S., & Cambria, E. (2020). Dialogue systems with audio context. *Neurocomputing*, *388*, 102–109.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833). Springer.

Zhao, X., Huang, W., Huang, X., Robu, V., & Flynn, D. (2021). Baylime: Bayesian local interpretable model-agnostic explanations. In *Uncertainty in artificial intelligence* (pp. 887–896). PMLR.

Zhao, W., Peng, H., Eger, S., Cambria, E., & Yang, M. (2019). Towards scalable and reliable capsule networks for challenging NLP applications. In *ACL* (pp. 1549–1559).

Zhou, Z.-H., Jiang, Y., & Chen, S.-F. (2003). Extracting symbolic rules from trained neural network ensembles. *AI Communications*, *16*(1), 3–15.