IEEE Intelligent Systems: Special Issue on AI Ethics and Trust
F. Chen, A. Holzinger, J. Zhou, K.R. Fleischmann, S. Stumpf; is6-23@computer.org

An operational framework for guiding human evaluation in Explainable and Trustworthy AI

Roberto Confalonieri
Department of Mathematics 'Tullio Levi-Civita', University of Padova, Italy, roberto.confalonieri@unipd.it

Jose M. Alonso-Moral
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Spain, josemaria.alonso.moral@usc.es

***Abstract*—The assessment of explanations by humans presents a significant challenge within the context of Explainable and Trustworthy AI. This is attributed not only to the absence of universal metrics and standardized evaluation methods, but also to complexities tied to devising user studies that assess the perceived human comprehensibility of these explanations. To address this gap, we introduce a survey-based methodology for guiding the human evaluation of explanations. This approach amalgamates leading practices from existing literature and is implemented as an operational framework. This framework assists researchers throughout the evaluation process, encompassing hypothesis formulation, online user study implementation and deployment, and analysis and interpretation of collected data. The application of this framework is exemplified through two practical user studies.**

TOPIC OF INTEREST Approaches for ensuring calibrated trust in AI; Fairness, accountability, and transparency in AI ethical framework implementations; Novel user experience design and evaluation methods for AI ethical framework implementations.

1. Introduction

The recent spread of AI systems, especially Deep Learning systems, which are deemed black boxes due to their lack of explainability, has raised the popularity of eXplainable AI (XAI) [1], [2], [3]. XAI research focuses on the development of methods and techniques that aim at providing explanations of how these hardly interpretable AI systems make decisions [4], [5]. The research on XAI is multidisciplinary, rooted in cognitive and social science, and strongly related to human-machine interaction, computational linguistics, etc. [6]. One of the most active research lines in XAI addresses the problem of human evaluation of explanations [7], [8], [9]. This problem is chal-

lenging not only because standard metrics and evaluation methodologies universally accepted do not exist, but also because designing a user study, which is typically how explanations are evaluated with humans in XAI, relies on a complex workflow. To be effective, user studies must be rigorously designed and implemented (regarding the explanation goals to be tested, the questions to be asked, the selection of participants and the evaluation of results).

To bridge this gap, we propose a methodology (and its implementation) for guiding human evaluation of explanations in XAI. The methodology is based on the taxonomy proposed by Chromik and Schuessler [9] that considers and aggregates several state-of-the-art taxonomies proposed in the literature [7], [10], [11], [12]. The methodology enacts this taxonomy through two phases commonly undertaken by researchers during the course of evaluation research: the planning stage and the execution and release stage, inclusive of the specific procedures within these phases. The methodology is implemented as an operational framework that aids AI researchers in the development of user studies for the evaluation of explanations. In particular, XAI researchers are guided in selecting the most suited modules depending on the use case considered. The framework automatically generates the code needed to produce a template for running an online questionnaire and collecting human responses. In the paper, we describe the methodology and its implementation through the development of two user studies, each consisting of three steps, namely the creation of an XAI questionnaire, the human evaluation of the XAI questionnaire created, and the analysis of the responses collected. In summary, the main contribution of this article is two-fold:

- A general methodology to guide human evaluation of automated explanations.
- An operational framework to assist AI researchers in developing user studies for human evaluation of explanations.

The rest of the document is structured as follows. Section 2 describes related work. Section 3 presents the methodology for designing user studies for the evaluation of explanations. Section 4 describes the implemented framework. Section 5 exemplifies, through a use case, how this framework can be used in practice. Section 6 concludes the document and outlines future work.

2. Related Work

Barredo Arrieta et al. [2] presented an overview of the literature and contributions in the XAI research field. They proposed and discussed a taxonomy of contributions related to the explainability of Machine Learning models. This survey is complementary to the historical perspective presented by Confalonieri et al. [3]. In addition, Barredo Arrieta et al. provided readers with a novel definition of XAI as "given an audience, an explainable AI is one that produces details or reasons to make its functioning clear or easy to understand." This definition considers some prior conceptual propositions described in the literature, with a significant focus on the target audience. As previously noted by Miller [6], tailoring explanations for the target audience is crucial for producing effective explanations. Moreover, different users may need different explanations, which should be adapted to the given task and context. Guidotti et al. [4] also provided readers with an overview of XAI techniques for explaining black-box AI systems, with especial emphasis on post-hoc approaches. It is worth noting that explanations are usually contrastive, so factual and counterfactual explanations are complementary when explaining AI systems [13], [14]. More recently, Ali et al. [1] provided a review of XAI techniques from four axes, namely data explainability, model explainability, post-hoc explainability, and assessment of explanations.

Rudin et al. [5] pointed out evaluation of explanations as a major challenge, among others, to be faced in the context of XAI. Accordingly, Hoffman et al. [7] introduced key concepts for measuring the quality of an XAI system—including the quality of explanations—derived from the integration of extensive research literature and psychometric assessments. The authors proposed a conceptual model of the explaining process. According to this model, initial instructions in how to use an AI system enable a user to form an initial mental model of the task to

be solved and the AI system.[1] Then, subsequent interactions with the system through, for instance, system-generated explanations, allow the participants to refine their mental model about the way in which the system works. The underlying assumption of this model is that when the explanations provided to users are of high quality and meet their expectations, these explanations will enable users to acquire deeper insights into the internal mechanisms of the system. As a result, their comprehension of the system improves, leading to further refinement of their initial mental model. In addition to the conceptual model definition, Hoffman et al. proposed methods that aim to evaluate the goodness of explanations.

Vilone and Longo [16] aggregated scientific studies that classify theories and notions related to the concept of explainability as well as the evaluation approaches for XAI methods via a hierarchical system. Such system is built on an analysis of existing taxonomies and peer-reviewed scientific material. The literature review highlighted various notions and requirements that an explanation should meet to be understood by the end users and provide actionable information for decision-making. The authors also described methods to evaluate to what degree the explanations generated by an AI system meet the evaluation requirements. In their analysis, they discovered a lack of consensus among scholars on how an explanation should be defined and validated. To this end, they proposed to incorporate 35 notions related to explanations— including algorithmic transparency, actionability, causality, to name a few—in the definition of explainability. The aim of incorporating these notions was to cover a broad set of attributes and dimensions of explainability, applicable in various contexts and application domains.

Furthermore, Cromik and Schuessler [9] proposed a taxonomy iterated through a systematic literature review to understand better how researchers across various disciplines approach human evaluation of explanations related to black-box AI systems. The proposal came after identifying a lack of consensus among the involved disciplines regarding the assessment of effectiveness in automated explanations, especially when humans are involved. Considering a Human-Computer Interaction perspective, the authors evaluated the scholars' study design for different explanation goals. They grouped the relevant dimensions for the evaluation of explanations with human subjects into task-related, participant-related and study design-related dimensions. Their taxonomy guides researchers and practitioners in designing and executing user studies for the evaluation of explanations.

The work by Van der Lee et al. [8] provided a review of how human evaluation is accomplished in Natural Language Generation (NLG) systems. They recommended a set of best practices that are related to the phases researchers typically go through when conducting an evaluation research, namely planning stage, execution and release stage, along with the precise procedures within these phases. Whilst their proposal aimed to contribute to the quality and consistency of human evaluation in the NLG domain, the proposed phases are also of interest in the context of XAI. They indeed provide the basis for specifying a methodology for the engineering of human evaluation studies.

In addition, Shin et al. [17], [18] addressed the challenging problem of algorithmic explainability from a human factors' perspective. In [17], the authors explored how explainability in AI affects user trust and attitudes, focusing on causability as a precursor to explainability. Their results reveal the dual roles of causability and explainability in building trust and emotional confidence, offering implications for enhancing trust through the inclusion of these aspects in AI systems for transparent decision-making processes. The study in [18] investigated user perceptions of fairness and transparency in AI systems, particularly within over-the-top (OTT) platforms. Through a mixed-method approach, authors explore how normative values influence user sense-making processes and how a composite concept of 'transparent fairness' affects perceived quality and credibility, ultimately proposing a theoretical model that positions transparent fairness as a vital attribute for trustworthy algorithmic media platforms. This fact is well aligned with the

[1]For user, we refer to the participant of a user study, and for mental model, we refer to the user understanding of the AI system whose explanations are being evaluated. This is in agreement with the evaluation protocol supported by psychological models of explanations as defined by DARPA [15].

Ethics Guidelines for Trustworthy AI published by the European Commission, but also with the European AI Act[2].

The above review of previous studies reveals that research in human evaluation of automated explanations is scattered and there is a lack of consensus. Furthermore, evaluating the goodness and effectiveness of automated explanations is a prerequisite for ensuring calibrated trust in AI.[3] To this end, in this paper, we introduce a general methodology for guiding human evaluation in XAI, as a step forward towards the implementation and validation of a human-centric Trustworthy AI in agreement with ethical values.

3. Methodology

A methodology for human evaluation of explanations is proposed in Figure 1. This methodology is based on concepts and ideas taken from the approaches reviewed in the previous section.

The methodology consist of two stages, namely Planning and Execution, that are associated with the definition of evaluation research, as proposed in the guidelines of Van der Lee et al. [8]. Whilst the Planning stage focuses on *what* to evaluate, the Execution stage focuses on *how* the evaluation is carried out. To define the steps within each stage, we adopted several concepts from the taxonomy proposed by Chromik and Schuessler [9]. Namely, in the Planning stage, we consider their type of evaluation, types of explanation goals, and type of participants. In the Execution stage, we consider their types of questions. In this stage, we also foresee the creation of the user mental model through initial instructions on the domain of interest, on the way in which the explanations are to be interpreted, and the evaluation is conducted. The idea of creating a user mental model through initial instructions is borrowed from the conceptual model of the explanation process defined by Hoffman et al. [7].

In the following sections, we provide more details about these two stages and the specific steps within each of them.

Planning Stage

The planning stage includes three steps: (i) definition of the evaluation type; (ii) definition of the explanation goal(s); and (iii) definition of the target audience.

There are two types of evaluations that can be carried out [8], [16]:

- **Qualitative Evaluation** consists of open-ended questions, looking for general insights about the AI system under evaluation.
- **Quantitative Evaluation** consists of close-ended questions for testing hypotheses with statistical data analysis.

In the context of XAI, an explanation is usually provided to humans with the intent of meeting specific requirements or *explanation goals*, in terms of the hypothesis to be validated, which should be clearly stated a priori. We have adopted the following nine explanation goals from [9]:

- **Transparency** to explain how the AI system works.
- **Scrutability** to allow users to tell the AI system it is wrong.
- **Trust** to increase users' confidence in the AI system.
- **Persuasiveness** to convince users to perform actions.
- **Satisfaction** to increase the ease of use or enjoyment of users when interacting with an AI system.
- **Effectiveness** to help users make good decisions.
- **Efficiency** to help users to make decisions faster.
- **Education** to enable users to generalize and learn.
- **Debugging** to enable users to identify defects in the AI system.

The last step in the Planning stage deals with the definition of the *target audience*, that is, the recipients of the explanation(s) who will take part in the evaluation. The target audience can vary depending on the explanation goal(s) previously defined. In addition, it influences the recruiting method and the number of participants. When the end users of an AI system are considered, they

[2] https://artificialintelligenceact.eu/

[3] Calibrated trust in AI refers to the degree of confidence that a user places in an AI system. Calibrated trust is important because it helps users make informed decisions about when to rely on the AI's output and when to seek additional information or take alternative actions. This concept highlights the need for an AI system to provide its users with good (e.g., transparent and accurate) explanations, ensuring that the trust users place in the system is appropriately aligned with its actual performance.

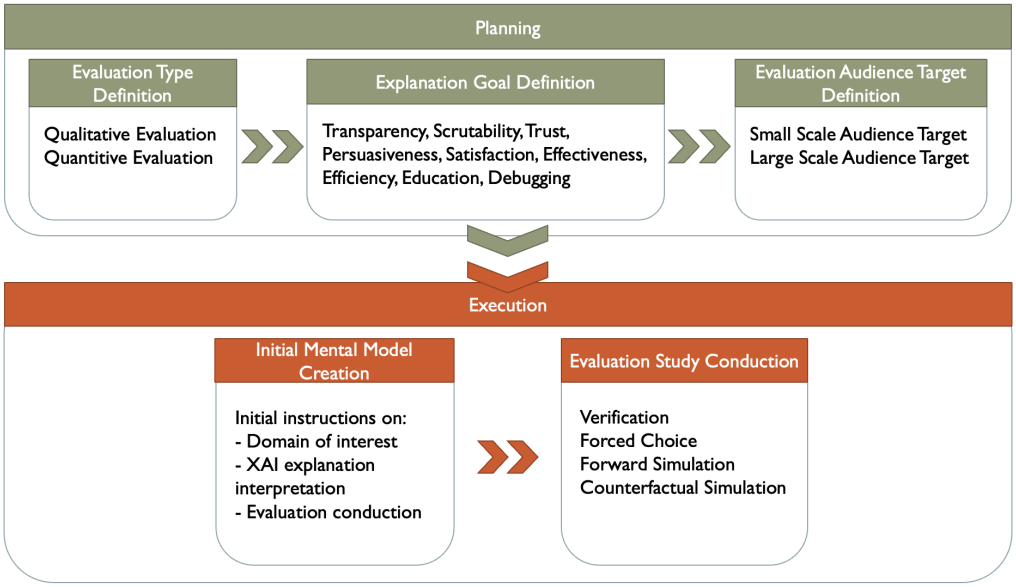

Figure 1. Overview of the methodology for human evaluation of explanations.

can be recruited in large numbers, for example, through a crowd-sourcing platform. Nevertheless, this approach cannot be followed if the evaluation requires domain experts, whose availability is usually much lower than in case of end users.

It is worth noting that Barredo Arrieta et al. [2] distinguished between different target audiences in XAI: users affected by model decisions, domain experts of the model (e.g., physicians or insurance agents), regulatory bodies or agencies, managers and executive board members, and data scientists or developers. For the sake of simplicity, we consider the two following groups:

- **Small Scale Target Audience**: data scientists, developers, regulatory bodies, managers or executive board members.
- **Large Scale Target Audience**: end users of an AI system.

Once the type of evaluation, the explanation goals, and the participants of the user study have been defined, the next stage pays attention to how the evaluation should be designed and carried out.

Execution Stage

This stage consists of: (i) the creation of an initial mental model for the user; and (ii) the running, as planned, of the evaluation study.

The first step is based on Hoffman et al.'s conceptual model of the XAI explanation process [7]. This process assumes that a good and satisfactory explanation is in agreement with a given user mental model. From the user point of view, having a good mental model means to trust the AI system and to show a good performance when using it. To build such a mental model, introductory information and instructions on how to carry out the evaluation task should be carefully prepared and provided to the participants in advance (e.g., information about the motivation, context and domain of interest, information about how the explanations should be interpreted and used, and information about how the evaluation is conducted). Accordingly, initial instructions aim to align the participant mental model with the knowledge required to participate in the evaluation effectively. Indeed, the initial instructions pave a common ground between the user mental model and the knowledge necessary to perform correctly the requested tasks.

The second step foresees the definition of user tasks, regarding all samples and stimuli to be involved in the experimental setting. Notice that asking participants of a user study to carry out certain tasks is one of the most common ways to assess the quality of explanations as already

pointed out by Doshi-Velez and Kim [10][4]. Moreover, in Chromik and Schuessler's taxonomy [9], the related material and methods are organized by considering the information provided to the participants and the information obtained in return. Here, we consider the following tasks:

- **Verification**: participants are provided with an input, an output, and an explanation. They are asked to rate their satisfaction with respect to the explanation provided (usually by means of a Likert-style scale).
- **Forced Choice**: participants are provided with an input, an output, and multiple explanations. They are asked to choose from multiple competing explanations, for example, by expressing an order between the explanations given or by pointing out the most suited one.
- **Forward Simulation**: participants are provided with an input (e.g., an instance data) and an explanation. They are asked to use the explanation to compute the output of the AI system whose explanations are being evaluated.[5]
- **Counterfactual Simulation**: participants are provided with an input, an output, alternative outputs (counterfactuals), and an explanation. They are asked to identify the changes to the input needed to obtain the given alternative outputs.

Prior to describing the operational framework that enacts the outlined methodology, let us enumerate all the steps that need to be accounted for in an evaluation pipeline. This enumeration is intended to ensure the reproducibility of the process:

- Declaring the research questions and related hypotheses to validate.
- Pre-registering the evaluation plan. This means setting up and documenting in advance all material and methods needed for evaluating the AI system under study. For example, it is important to distinguish between independent and dependent variables when defining samples and stimuli. It is also important to declare in advance the data management plan as well as the statistical analysis plan.
- Getting approval of the Ethics Committee in your institution, regarding issues such as recruitment and payment of subjects as well as data storing, processing and privacy.
- Implementing and conducting the study.
- Running statistical data analysis and reporting results which should validate (or not) the hypotheses that were stated at the beginning.

The last two steps are facilitated with the assistance of the framework to be introduced in the next section.

4. Implementing the Methodology

The methodology that we introduced in the previous section is a general scaffold that can be instantiated to meet different requirements and to implement different user studies. To instantiate this methodology, a Python toolkit was developed. This implements a pipeline that guides AI researchers[6], interested in the evaluation of explanations with humans, through running three software wizards: (i) the XAI Questionnaire Generation, (ii) the XAI Questionnaire Evaluation, and (iii) the XAI Questionnaire Analysis.

As we will see below, the wizards ease not only the implementation of a user study, but also its personalization for the specific use case at hand. The open source code developed for the XAI Questionnaire Generator[7] and the XAI Questionnaire Analyzer[8] is available at Github. The XAI Questionnaire Evaluation is, in turn, generated dynamically after completion of the XAI Questionnaire Generator. An illustrative use case, which exemplifies how the methodology is used in practice, will be presented in Section 5.

XAI Questionnaire Generation

The XAI questionnaire generator wizard automatically creates an XAI questionnaire template. This template guides the developer of the questionnaire in the creation of a user study. The template is generated by taking into account user

[4]The taxonomy of interpretability issues proposed by Doshi-Velez and Kim [10] was already adopted in some publications [19], [20].

[5]To explain the task further, let us imagine that the AI system is a neural network classifier and the explanation is presented as a decision tree (DT). Then, this task would imply to use a DT to classify a given instance.

[6]The toolkit assumes a certain acquaintance with Web programming and software development.

[7]https://github.com/marcozenere/XAI_Survey_Generator

[8]https://github.com/marcozenere/XAI_Survey_Analyser

requirements which are gathered by means of three questions. Each question is associated with one of the steps that were defined in the planning stage of the methodology:

1) What type of evaluation would you like to perform? The possible answer is *qualitative evaluation* or *quantitative evaluation*.
2) What type of user task would you like to select considering the explanation goal(s)? The possible answer is *Verification*, *Forced Choice*, *Forward Simulation*, or *Counterfactual Simulation*.
3) How many questions would you like to have in your questionnaire? The possible answer is an integer value higher or equal than one.

This wizard was implemented as a Web application through a Jupyter notebook. This notebook is structured into three units, one for each of the questions presented above. Each page contains all the necessary information for understanding the given questions, and the options to choose from.

Notice that, the notebook is meant to be followed linearly, from Question 1 to Question 3. However, the notebook allows the user to go back to the previous sections and revise the answers if desired. Once the user is satisfied with the design of the questionnaire, a questionnaire template is automatically generated. The template comes in the form of a Jupyter notebook itself, a readme file, and a text file for running it in *Binder*.[9] The notebook was implemented using the *ipywidgets*[10] and *nbformat*[11] libraries. The first library consists of a set of widgets to create the options in each question (e.g., radio buttons and text widgets). The second library includes methods that are needed to generate a Jupyter notebook. The interested reader can find further implementation details in the associated supplementary materials.

XAI Questionnaire Evaluation

Before running the user study, the AI researcher has to fill in and customize the questionnaire template previously generated. The template contains the necessary functions to create an XAI questionnaire, such as basic checks and placeholders that need to be completed with the information of the specific use case at hand. The template consists of seven distinct sections. These sections take inspiration from the conceptual model for XAI evaluation proposed by Hoffman et al. [7]. The first four sections aim to create a good user mental model through initial instructions and examples. Sections 5 and 6 correspond to the actual XAI assessment. Section 7 aims to gather information about the participants for the demographic analysis. In detail, the questionnaire template consists of the following sections:

1) The **Welcome Section** summarizes the research goal and who is in charge of running the questionnaire, as well as GDPR[12] information and the definition of the terms for participating in the evaluation.
2) The **Introductory Section** provides information about the domain of interest and about the explanations generated by the AI system under study.
3) The **Example Section** provides a detailed example of what participants will be asked to do in the evaluation. The aim is to let participants be familiar with the evaluation task, and create a good mental model.
4) The **Comprehension Section** aims to evaluate the user mental model. This evaluation is usually conducted by asking participants to carry out a certain task and by measuring task related performance.[13]
5) The **Questionnaire Instructions Section** provides information of how the evaluation is conducted and instructions on how to complete the questionnaire.
6) The **Questionnaire Questions Section** includes the material for the actual evaluation.
7) The **Questionnaire Participant Information Section** comprises demographic questions about the participants, such as gender, age, education or English level.

[9] *Binder* is a service to deploy a Jupyter notebook via a Web link, https://mybinder.org

[10] https://ipywidgets.readthedocs.io/en/stable/

[11] https://nbformat.readthedocs.io/en/latest/

[12] GDPR stands for General Data Protection Regulation: https://eur-lex.europa.eu/EN/legal-content/summary/general-data-protection-regulation-gdpr.html

[13] For example, a participant might be first explained how to classify data instances and then asked for classifying one or more particular instances. Then, the accuracy of the provided answers can be used as a proxy to determine whether the participant was able to create a good mental model or not.

It is worth noting that the questionnaire template runs in a Jupyter notebook. This facilitates the development of code within a web-based interactive environment.

XAI Questionnaire Analysis

This stage deals with the analysis of the results collected during the previous evaluation with human subjects. Highlighting the key characteristics of the assessment in a human-comprehensible form can help in verifying the achievement of the research goal(s) defined at the beginning of the study. The analysis produces insights for the comprehension section (i.e., user mental model analysis), the questions section (i.e., XAI evaluation analysis), and the participant information (i.e., demographic analysis).

The XAI questionnaire analyzer computes the participant prediction accuracy and the time taken to answer a question when a specific explanation was displayed. This software package implements basic statistics (e.g., the total number of answers or the correct ones for each explanation presented). It also generates graphics summarizing these statistics, and the demographic information.

5. Illustrative Use Case

As a proof of concept, two user studies about the evaluation of explanations were developed. Following the methodology that we introduced in previous sections, the interpretability of explanations can be measured through human-grounded metrics, such as accuracy, time of response, and user-reported understandability. Accordingly, the evaluation can focus on the perceived interpretability of explanations rather than on their mechanistic creation.

In the following sections, we delve into how the introduced operational framework was used for the design and execution of the two user studies:

- The first user study aims at illustrating the use of the framework in a practical use case (regarding the goodness of explanations under consideration).
- The second user study aims at evaluating the satisfaction of AI researchers with the framework (regarding understandability, confidence and predictability issues).

Further details about the design of these user studies and the analysis of the results collected can be found in the supplementary material accompanying the paper.

Material

In both studies, we considered the wine dataset[14] for generating the models and explanations to be evaluated. The wine dataset consists of 178 data instances (without missing values) related to a chemical analysis of wines grown in the same region in Italy but derived from 3 different cultivars (i.e., each instance is classified in one of the 3 given classes, each class is associated to a wine type). The dataset includes 13 numerical attributes which are the constituents of the wine determined after analyzing the raw data collected, as follows: alcohol, malic acid, ash, alcalinity of ash, magnesium, total phenols, flavanoids, nonflavanoid phenols, proanthocyanins, color intensity, hue, OD280/OD315 of diluted wines, and proline.

The decision tree classifier algorithm of sklearn with 10-fold cross-validation was used on the wine dataset to train two decision trees (DTs) with different complexities/sizes: a 3-layers DT representation (including 7 decision nodes and 4 leaf nodes, as depicted in Figure 2) and a 5-layers DT representation (including 11 decision nodes and 4 leaf nodes). They achieved a prediction accuracy of 92.7% and 97.8%, respectively, in agreement with the well-known interpretability-accuracy tradeoff in which a more complex model, without overfitting, is expected to achieve a higher accuracy. The graphical representation of the DTs were taken as a *proxy explanation* of the underling classification task. Thus, in the rest of this section, Explanation 1 is associated with the 3-layers DT while Explanation 2 is associated with the 5-layers DT.

The visual representation of the DTs was produced using the *dtreeviz* library[15]: the root and internal nodes are represented as histograms and the leaf nodes are depicted as pie charts. On the one hand, each histogram highlights the class distribution w.r.t. the considered attribute and depicts the threshold value used for the binary splitting. On the other hand, each pie chart highlights the dominant class in a leaf node.

[14] https://archive.ics.uci.edu/ml/datasets/wine
[15] https://github.com/parrt/dtreeviz

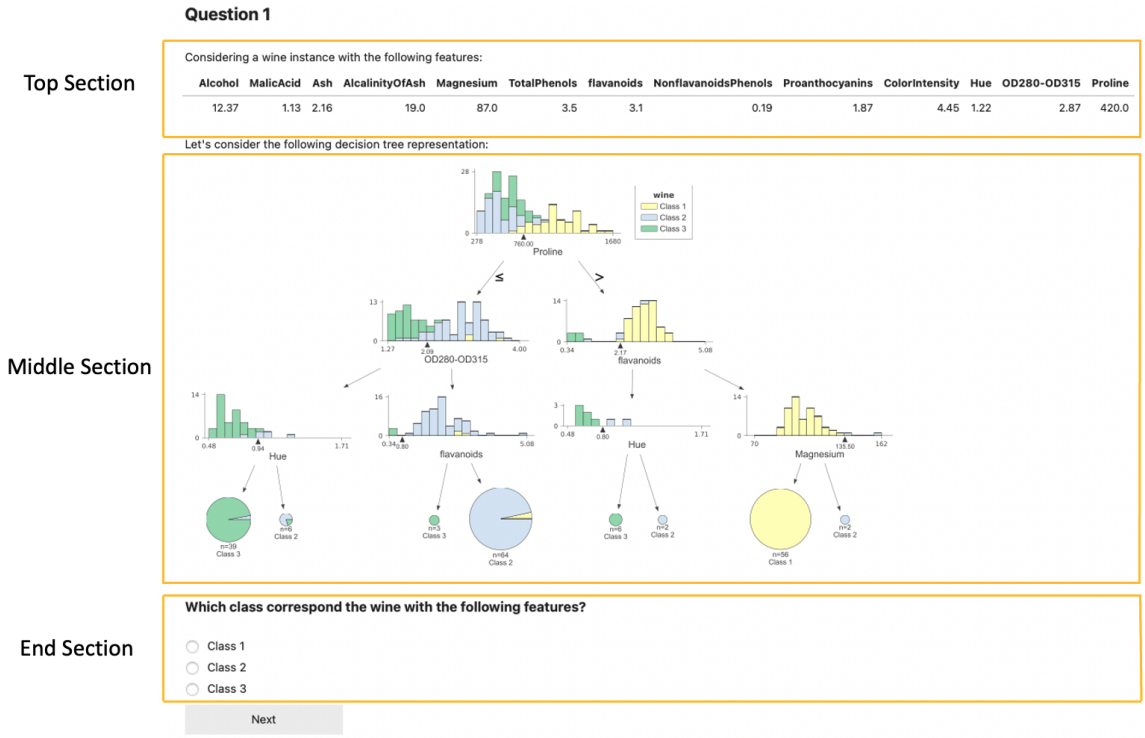

Figure 2. "Questionnaire Questions" section in the implemented user studies.

Methodology

According to the proposed methodology, a user study comprises two distinct stages, i.e., the planning and execution stages.

On the one hand, in the planning stage, the AI researcher has to make decisions about the evaluation type, the explanation goals, and the participants. The goal of the first user study was to evaluate the interpretability of explanations modeled as DTs. To this end, we opted for the use of close-ended questions to facilitate the statistical analysis of results. In addition, since evaluating the interpretability of explanations can be seen as a way to measure the transparency of an AI system [10], we selected *transparency* among the nine possible explanation goals. Finally, a small scale target audience was identified as the most suitable for this study.

On the other hand, in the execution stage, we had to deal with both the creation of the user mental model and the evaluation of the explanations. As far as the creation of the user mental model is concerned, we opted for providing participants with information about the domain of interest, as well as the DT representation, its use as a model to classify a given instance, and the experiment conduction via initial instructions. For the evaluation of explanations, we considered the forward simulation task as it is a task that is typically used to evaluate the transparency of an AI system [9]. It is worth noting that the task-based questions were displayed as in a randomized control test to prevent bias in the collected results. Considering these requirements, the execution stage was designed as follows:

- Among the user tasks, the "Forward Simulation" task was selected.
- Stimuli were "randomly displayed" to participants in the experiment.
- To judge the participant knowledge and understanding, the correctness of the answers given throughout the evaluation was computed.

Based on this information we implemented a questionnaire using the wizard previously described. We used the "XAI Questionnaire Generator" tool with the following options: the "quantitative evaluation" as evaluation type, "Forward Simulation" as type of user task, and 5 as the number of questions. Then, the generated ques-

tionnaire was the ground for the "XAI Questionnaire Evaluation" stage where we edited and customized the questionnaire template with the aforementioned requirements. In the first sections of the questionnaire, we inserted the initial instructions for the participants: (i) a brief description of the wine dataset, (ii) an explanation of how to interpret the given DT representation, and (iii) a step by step explained example along with the outline of the rest of the questionnaire. Then, we added content to the "Comprehension" section, with two close-ended questions which were aimed for building up, and measuring the goodness of, the user mental model in terms of the correctness of the related answers. Namely, we asked each participant to guess the class to which the given data instance belongs to, regarding the path that is activated in the DT. Then, the "Questionnaire Questions" section (see Figure 2) was completed following the predefined template for "Forward Simulation" user tasks. The last section of the questionnaire, the 'Questionnaire Participant Information', was kept as it was in the template.

In the second study, the questionnaire followed the same structure as the one just described for the first study. However, in this case the "Questionnaire Questions" section included only four task-oriented questions (a subset of those considered in the first study; with two questions related to Explanation 1 and two questions related to Explanation 2). In addition, we explained participants how our operational framework works with a brief presentation (covering all related steps in the Planning and Execution stages: from generating the questionnaires until analyzing the collected results). Then, we added at the end three general questions about the operational framework (regarding understandability, confidence and predictability issues).

The implemented questionnaires are with the supplementary material accompanying the paper.

Participants

The first user study involved a relative small sample of 13 participants. They were mainly master students in Computer Science who participated in the user study as volunteers. Most of them were male (84.6%), between 21 and 29 years old (76.9%), with graduate-level education (84.6%) and English proficiency (76.9%).

The second study involved a broader sample of participants (47) who took part in a series of 3 independent sessions (each session took 1 hour). The target audience included students with technical background who were enrolled in training courses related to XAI. Only those 38 participants (73.7% of them were male and graduate, 89.5% with English B2 or higher, all between 21 and 29 years old) who passed the comprehension test went on with the rest of the study.

Results

In the last step of the execution stage, the "XAI Questionnaire Analyzer" assisted us in the analysis of the collected data and gave insights into the following sections of the questionnaire: "Comprehension", "Questions", and "Participant Information".

In both studies, the "Comprehension" questions asked to participants were:

- **Q1**: Which class corresponds to the wine with the following features?
- **Q2**: Which of the following features did you consider for the classification task?

In the first study, 91.7% of the participants answered Q1 correctly, and all of them answered correctly to Q2. Considering these results, we could conclude that the information provided through initial instructions was sufficiently good to create an appropriate user mental model.

In the "Questions" section, we compared Explanation 1 (3-layers) and Explanation 2 (5-layers) in terms of prediction accuracy of the participants and the time taken to answer each question. Each participant was shown five questions about Explanation 1 and Explanation 2. The type of explanation shown to each participant was randomly chosen when the questionnaire was started. In general, Explanation 2 yielded a higher level of prediction accuracy compared to Explanation 1. However, when Explanation 2 was presented, participants took longer to answer than in case of Explanation 1; what is in agreement with the fact that the second tree is structurally more complex than the first one. In detail, the time (in seconds) taken by the participants was:

- **Explanation 1**: Mean = 30.0s, Median = 20.324s, Standard Deviation = 20.797s.

- **Explanation 2**: Mean = 41.164s, Median = 30.876s, Standard Deviation = 34.599s.

The Null hypothesis (H0: "The medians of the differences between the two group samples are equal") was rejected (p-value=0.043) when running the Wilcoxon non-parametric test.

The data considered above was well balanced among the participants: Explanation 1 was displayed thirty-one times, whereas Explanation 2 was displayed thirty-four times. For more details about the results, we refer to the supplementary material accompanying the paper.

Results in the second user study, regarding the comparison of accuracy between decisions assisted by the two types of explanations, were in agreement with those already discussed for the first study (i.e., most participants were able to accomplish well the required tasks no matter the complexity of the given explanation). Considering these results and the research goal of the evaluation in the use case, we could state that a longer but more accurate explanation (i.e., Explanation 2) is preferable. When Explanation 2 was displayed, participants understood the decision-making process of the AI classifier better than when Explanation 1 was shown. As a result, Explanation 2 enabled the construction of a better user mental model than Explanation 1, which led to greater confidence in the AI system and better performance when using it. It is worth noting that this result seems to contradict somehow the believe in the so-called interpretability-accuracy trade-off. We can expect that a simpler model (3-layers) is easier to process than a more complex one (5-layers). However, a more compact system is not necessarily easier to understand because transparency and understandability are not always well correlated when dealing with human evaluation, where user background and context matter. This is an empirical example that, quoting to Rudin et al. [5], 'there is no scientific evidence for a general tradeoff between accuracy and interpretability'.

Finally, regarding answers to the 3 general questions at the end of the second study, we can conclude that most participants were satisfied with the proposed framework:

- 86.8% of participants agree (or strongly agree) with the statement "The provided explanations helped me understand how decision trees work".
- 65.8% of participants agree (or strongly agree) with the statement "I am confident in the tool. I think it works well".
- 65.8% of participants agree (or strongly agree) with the statement "The outputs of the tool are predictable".

6. Concluding Remarks

In this paper, we have introduced a novel methodology for addressing the challenge of evaluating the goodness of explanations in XAI, what we consider as a prerequisite for ensuring calibrated trust in AI. This methodology turns up as an effort to put together the best practices found in the literature. It is fully operational as shown in an illustrative use case. Moreover, it is released as an operational framework for filling the gap between theory and practice when assisting AI researchers in the entire evaluation process, from declaring the hypotheses to validate until analyzing the collected data and discussing the reported results, passing by the elaboration and deployment of online questionnaires.

Due to the infeasibility of addressing all conceivable evaluation scenarios, the use case presented served as a mere illustration of potential analyses. In the interest of ensuring both comprehensiveness and replicability, we released a well-documented software which includes a core with basic functions which are easy to customize and extend for different user studies. Accordingly, as a future work, we plan to apply our methodology to other use cases, updating the related software with new functionalities.

Acknowledgments

The authors would like to thank Marzo Zenere for the implementation of the Python wizard during his MSc thesis. This work is supported by MCIN/AEI/10.13039/501100011033 (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 which are co-funded by the ERDF/FEDER program).

REFERENCES

1. 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*, vol. 99, p. 101805, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253523001148
2. A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253519308103
3. R. Confalonieri, L. Coba, B. Wagner, and T. R. Besold, "A historical perspective of explainable artificial intelligence," *WIREs Data Mining and Knowledge Discovery*, vol. 11, no. 1, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1391
4. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.
5. C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable machine learning: Fundamental principles and 10 grand challenges," *Statistics Surveys*, vol. 16, no. none, pp. 1 – 85, 2022.
6. T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019, https://doi.org/10.1016/j.artint.2018.07.007.
7. R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable AI: Challenges and prospects," *ArXiv*, vol. abs/1812.04608, 2018.
8. C. van der Lee, A. Gatt, E. van Miltenburg, and E. Krahmer, "Human evaluation of automatically generated text: Current trends and best practice guidelines," *Computer Speech & Language*, vol. 67, p. 101151, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S088523082030084X
9. M. Chromik and M. Schuessler, "A taxonomy for human subject evaluation of black-box explanations in XAI," in *ExSS-ATEC@IUI*, 2020.
10. F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv: Machine Learning*, 2017.
11. I. Nunes and D. Jannach, "A systematic review and taxonomy of explanations in decision support and recommender systems," *User Modeling and User-Adapted Interaction*, vol. 27, pp. 393–444, 2017.
12. N. Tintarev and J. Masthoff, "A survey of explanations in recommender systems," *2007 IEEE 23rd International Conference on Data Engineering Workshop*, pp. 801–810, 2007.
13. 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*, vol. 9, pp. 11 974–12 001, 2021, http://dx.doi.org/10.1109/ACCESS.2021.3051315.
14. R. Guidotti, "Counterfactual explanations and how to find them: literature review and benchmarking," *Data Mining and Knowledge Discovery*, pp. 1–55, 2022.
15. D. Gunning, E. Vorm, J. Y. Wang, and M. Turek, "DARPA's explainable AI (XAI) program: A retrospective," *Applied AI Letters*, vol. 2, no. 4, p. e61, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ail2.61
16. G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Information Fusion*, vol. 76, pp. 89–106, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253521001093
17. D. Shin, "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai," *International Journal of Human-Computer Studies*, vol. 146, p. 102551, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071581920301531
18. D. Shin, J. Lim, N. Ahmad, and M. Ibahrine, "Understanding user sensemaking in fairness and transparency in algorithms: algorithmic sensemaking in over-the-top platform," *AI and Society*, 2022.
19. R. Confalonieri, T. Weyde, T. R. Besold, and F. M. del Prado Martín, "Using ontologies to enhance human understandability of global post-hoc explanations of black-box models," *Artificial Intelligence*, vol. 296, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370221000229
20. G. Righetti, D. Porello, and R. Confalonieri, "Evaluating the interpretability of threshold operators," in *Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management (EKAW 2022)*, ser. LNCS, vol. 13514. Springer, 2022, pp. 136–151. [Online]. Available: https://doi.org/10.1007/978-3-031-17105-5_10

Roberto Confalonieri received his Ph.D. degree in Artificial Intelligence at the Technical University of Catalonia, Spain, in 2011. He is Associate Professor of Computer Science at the Department of Mathematics "Tullio Levi-Civita" at the University of Padua, Italy. He is Senior Editor of the Cognitive Science Research journal (Elsevier) and Associate Editor of the Neurosymbolic Artificial Intelligence journal (IOS Press).

Jose Maria Alonso-Moral received his Ph.D. degree in Telecommunication Engineering at the Technical University of Madrid, Spain, 2007. He is Associate Professor at CiTIUS-USC, Chair of the Task Force on "Explainable Fuzzy Systems" in the Fuzzy Systems Technical Committee of the IEEE Computational Intelligence Society, Associate Editor of the IEEE Computational Intelligence Magazine, and coordinator of the H2020-MSCA-ITN-2019 project entitled "Interactive Natural Language Technology for Explainable Artificial Intelligence" (NL4XAI).