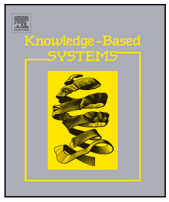
Knowledge-Based Systems 302 (2024) 112305

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



iSee: A case-based reasoning platform for the design of explanation experiences

Marta Caro-Martínez [a,*], Juan A. Recio-García [a], Belén Díaz-Agudo [a], Jesus M. Darias [a], Nirmalie Wiratunga [b], Kyle Martin [b], Anjana Wijekoon [b], Ikechukwu Nkisi-Orji [b], David Corsar [b], Preeja Pradeep [c], Derek Bridge [c], Anne Liret [d]

[a] *Facultad de Informática, Universidad Complutense de Madrid, Spain*
[b] *School of Computing, Robert Gordon University, Aberdeen, Scotland, United Kingdom*
[c] *Insight Centre for Data Analytics, School of Computer Science and IT, University College Cork, Cork, Ireland*
[d] *British Telecommunications, France*

ARTICLE INFO

Keywords:
eXplainable artificial intelligence
Trustworthy AI
Case-based reasoning

ABSTRACT

Explainable Artificial Intelligence (XAI) is an emerging field within Artificial Intelligence (AI) that has provided many methods that enable humans to understand and interpret the outcomes of AI systems. However, deciding on the best explanation approach for a given AI problem is currently a challenging decision-making task. This paper presents the *iSee* project, which aims to address some of the XAI challenges by providing a unifying platform where personalized explanation experiences are generated using Case-Based Reasoning. An explanation experience includes the proposed solution to a particular explainability problem and its corresponding evaluation, provided by the end user. The ultimate goal is to provide an open catalog of explanation experiences that can be transferred to other scenarios where trustworthy AI is required.

1. Introduction

Artificial Intelligence (AI) is integrated into our daily lives, whether it is suggesting movies through a recommender system, predicting the weather, or providing information through a chatbot. However, AI is increasingly finding application in more critical domains, such as predicting the presence of a tumor [1], making a decision in a trial [2], or detecting possible attacks to citizen security [3]. Consequently, understanding the working of an AI model becomes important. Accordingly, eXplainable Artificial Intelligence (XAI) plays a key role in ensuring that we can place trust in AI decisions when they impact crucial areas of our lives [4], for example when predicting illnesses such as gestational diabetes [5], acute myocardial infarction [6], or COVID-19 prognosis [7].

The XAI research field aims at providing explanations about an AI model's behavior and outcomes. The purposes of such explanations are often multi-faceted [8]. For example, explanations may aim to enhance task performance when using the AI [9], build user trust in the system [10], or uncover potential ethical and bias issues [11]. In safety-critical domains (such as medical, judicial, or security use cases), there are often legal requirements for traceability of algorithmic decision-making to provide accountability for autonomous outcomes [12]. Transparency is therefore not only a social responsibility [13], but increasingly required for compatibility with legislation [14].

Interest in XAI has been rising [15]. As a result, there is currently a wide range of explanation methods on offer, which we hereafter refer to as explainers, that can be used to provide explanations in different contexts and according to different needs, such as the users' goals, the scope of the desired explanation, or the data that is available [16]. Although this is good news for both research and industry, deciding on the best explanation approach for a given AI context and user is currently a challenging decision-making task [17].

The goal of the iSee project,[1] whose motivation is presented in Section 2, is to overcome this problem by providing a unifying platform that allows AI developers to reuse and apply the most satisfactory explanation strategy for a given AI scenario [18,19]. Despite significant advancements in explainable AI, notable research gaps remain. for example, current XAI models often lack mechanisms for interactive feedback, limiting their ability to single shot explanations. Additionally, existing systems frequently fail to adequately address the diverse and nuanced intents of users, resulting in explanations that may not align with individual needs or situational demands. These gaps highlight the

* Corresponding author.
E-mail address: martcaro@ucm.es (M. Caro-Martínez).

[1] https://isee4xai.com/.

https://doi.org/10.1016/j.knosys.2024.112305
Received 6 May 2024; Received in revised form 25 July 2024; Accepted 31 July 2024
Available online 8 August 2024

need for more user-centered approaches in XAI research, focusing on enhancing interactivity and contextual sensitivity to better meet user expectations and requirements.

To achieve this, the iSee platform has methods for assessing the effectiveness of an explanation strategy for the given AI scenario, along with knowledge structures that can precisely describe the factors that constitute the user context. These knowledge structures facilitate systematic comparison of descriptions of user contexts in order to identify similar scenarios and understand differences in context to manage adaptation to variations in user explanation needs.

The explanation strategies are collected through the development of several real-world use cases. In this way, we generate a catalog of strategies that can be transferred to other AI solutions. From this catalog, iSee applies a Case-Based Reasoning (CBR) [20] methodology to retrieve from its catalog the most suitable explanation strategy that it has for the given domain and task, and then adapts it to meet the end-user's requirements. This CBR approach, presented in Section 3, fits perfectly with our goal of providing the research community and industry with a self-learning catalog of reusable explanation experiences. First, it allows AI developers (called design-users in iSee) when they require an explainability solution to define their needs and eventually obtain the most suitable explanation strategy. After these design-users have applied the strategy to their own situation, they can, if they wish, agree to share their own explanation experiences with the platform. Secondly, XAI researchers have the opportunity to contribute new explainers to the catalog, thereby creating a pathway to research impact. This approach not only adheres to best practices in fostering collaboration and knowledge-sharing, but it also ensures the experiential knowledge available on the platform grows over time. Consequently, other users can benefit from the reuse of these vetted explanation experiences.

To the best of our knowledge, up to this date, no other XAI-oriented platform focuses on providing fully personalized and user-centered experiences for explanations. While there is a broad selection of explainers in XAI-specific libraries such as Alibi [21], Dalex [22], and Xplique [23], the libraries typically act as just a compendium of different approaches aimed at developers, neglecting the user experience aspect [24]. There are, however, a few examples of user-centered frameworks and tools for explainability, such as the work by Retzlaff et al. [25]; the Explainer Dashboard [26]; and the Dalex Arena. The framework by Retzlaff et al. is a set of decision support guidelines modeled as a decision tree to choose the best explainer, from among seven of the most popular ones, to apply to a use case. While this model is a good starting point, the iSee platform takes into account many more features regarding the AI model, type of user and other explanation requirements. It can also recommend complex XAI strategies that might be a combination of the explainers in its catalog (currently 67 of them), instead of recommending only one single explainer. The Explainer Dashboard and the Dalex Arena are tools that aim to improve user experience by simultaneously displaying the results of different explainers and allowing the user to test a variety of pre-computed explanations through a user-friendly interface. While the iSee platform provides similar functionality, where users can build their explanation strategies and execute explainers on-demand, it goes one step further by leveraging the advantages of CBR to recommend the best strategies based on the user profile and the characteristics of the given use case. In addition, as opposed to the previously cited tools, iSee provides a broad catalog of explainers that can be applied to machine learning models trained on different data types, such as tabular, text, images, and time-series data. While this catalog contains well-known explainers, several explainers have been specifically developed by the iSee team members and collaborators. Among these explainers, we highlight instance-based methods such as DiSCERN [27] and PertCF [28], IREX [29], and time-series explainers based on CBR for forecasting such as iGenCBR [30] and CBRFox [31].

The iSee platform is described in Section 4. It provides an explanation strategy evaluation cockpit that can be integrated into any proprietary software ecosystem to provide explainability. This cockpit allows design-users to define the contextual requirements of their explanation needs (including the domain, dataset, and model to be explained, as well as the different stakeholder groups who use the system). This information is used to recommend an explanation strategy comprising one or more explainers to provide coverage of all the contextual requirements. The platform then supports design-users to deploy this explanation strategy using a chatbot, creating an opportunity for end-users to interact with the different explainer algorithms within the strategy and explore all aspects of their explanation needs. End-users can then provide feedback, which is aggregated and visualized by the iSee platform, to inform the design-user as to how to refine the original explanation strategy to create a better explanation experience. All aspects of the experience (the contextual requirements supplied by the design-user, the recommended explanation strategy, and end-user feedback) are then stored as a case to support reuse in the future.

In some situations, the iSee platform may present ethical challenges for different stakeholder groups. In Section 6, we present an overview of the central ethical dimensions that the platform is likely to encounter. In Section 7, we will discuss our work, the current state of our project, and the expectations about the iSee platform.

2. Motivation

A right to obtain an explanation of the decision reached by a machine learning (ML) model is now a European Union (EU) regulation. Different stakeholders (e.g. managers, developers, auditors, etc.) may have different background knowledge, competencies, and goals, thus requiring different kinds of interpretations and explanations. Fortunately, there is a growing armory of ways of interpreting ML models and explaining their predictions, recommendations, and diagnoses. We will refer to these collectively as *explanation strategies*. As these explanation strategies mature, practitioners will gain experience that helps them know which strategies to deploy in different circumstances. What is lacking, and what iSee addresses, is the science and technology for capturing, sharing and re-using explanation strategies based on similar user experiences, along with a much-needed route to XAI compliance. Its vision is to improve every user's experience of AI, by harnessing experiences of best practices in XAI by users for users.

iSee facilitates XAI to the extent that users of AI can capture, share, and re-use their explanation experiences with other users who have similar explanation needs. An explanation experience captures information about the use of an explanation strategy. The strategies can be foundational ones, of the kind found in the research literature. However, user needs are often multi-faceted, and real-world applications and different users might require composite strategies, formed from combinations of the basic building blocks provided by one or more of the foundational strategies. The hypothesis proposed by iSee is that an end-user's explanation experience (like many other problem-solving experiences) must contain implicit knowledge that was required to solve their explanation need, such as the preferred strategy (foundational or composite) and, in the case of composites, the manner of combination. This way, iSee provides the necessary platform to capture experiences by enabling users to interact with, experiment with, and evaluate explanations. Experiences once captured can be reused, on the premise that similar user needs can be met with similar explanation strategies. Reuse may reinforce the usefulness of certain strategies for certain circumstances, and it can also expose situations where a suitable strategy has yet to be discovered.

There are already several libraries of explanation methods, including: Seldon Technologies Alibi Library; Sicara's tf-explain; Facebook's Captum; and IBM Research's AI Explainability 360. But none come with the kinds of representations of applicability and suitability conditions that iSee provides. Additionally, many XAI evaluation instruments have

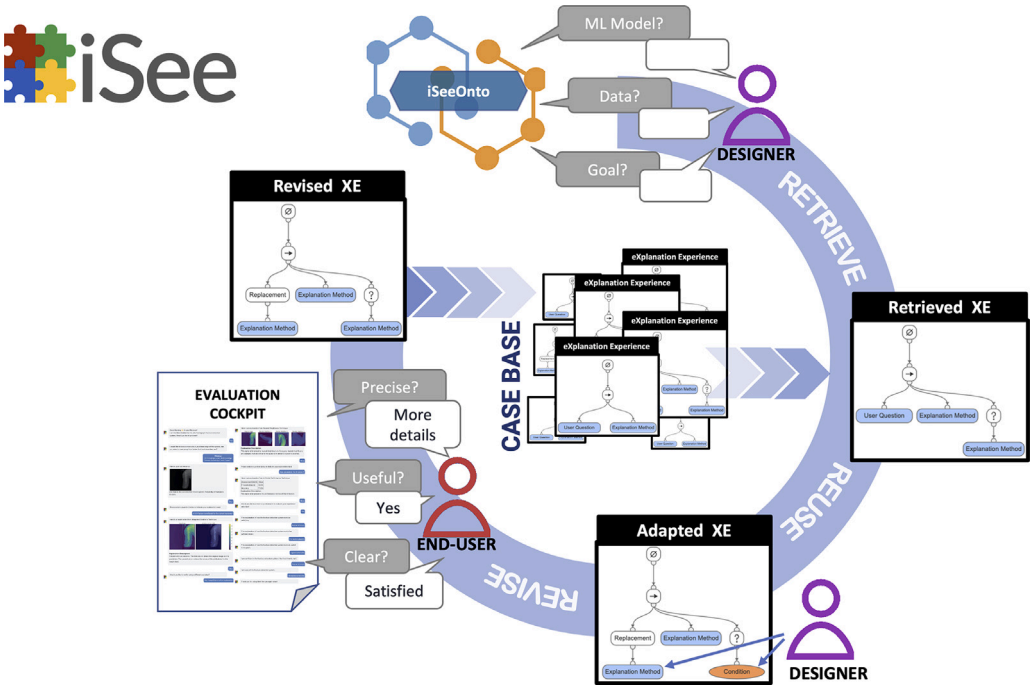


Fig. 1. The iSee CBR methodology.

been proposed. For example, there are published reviews, e.g. [32], and there are several recent evaluation frameworks, e.g. [33,34]. Only some of these address the variability of different explainers. Moreover, these frameworks are either extremely high level (giving general guidance but typically no technical implementation to support a user) or they are designed for a specific explainer or a specific domain, in which case they often lack guidance on transferability to alternative explainer methods or domains. Considering that the majority of these evaluation instruments lack good tool support, they tend to be demanding in both time and money.

The challenge is to provide a platform for the digital delivery of evaluation instruments by enabling in-the-wild experimentation that can reason about explanation suitability while observing the user on the job. The reasoning requires the integration of multiple user actions (from dialog through to interface interactions) before, during, and after explanations are presented. Hence, iSee employs a user-participatory approach. This promotes ongoing involvement and drives an open-science agenda since it involves collecting knowledge about XAI, shaped by users' experiences [35]. User research has provided insights into users' mental models (e.g., their beliefs, desires, goals, and intentions), including what they want to do with the iSee platform, how the system might be integrated into their ML ecosystem, when and how they will use it, and importantly it enables us to co-create applicability and suitability criteria for the explanation strategies. Participatory ontology discovery is known to enhance the quality of results [36] and is necessary for iSee, given the expected incompleteness of the existing literature and domain knowledge.

A major contribution of iSee is the *evaluation cockpit*, an environment that supports design-users who are seeking to deploy ML and XAI for their end-users. The cockpit offers to the design-users instruments to evaluate acceptability and suitability, and the facilities to set up their own experiments to evaluate explanation strategies. Interactions here are a proxy for a user's explanation experience, context, needs, and satisfaction, providing us with insights into their mental models [32]. The cockpit facilitates co-creation activities, addressing crucial questions such as what users expect given their varied roles and how they are to be represented in the CBR cycle.

3. The iSee CBR methodology

The CBR methodology [37] allows us to transfer the solutions of previous explanation experiences to suit the needs of new scenarios. Moreover, users can personalize solutions according to their preferences. The process of determining a suitable explanation method, in particular for black-box models, is a demanding and mainly knowledge-intensive task performed by AI engineers and designers. The iSee CBR methodology proposes storing the valuable experiences gained by experts and determining the right explanation method for their current use case. Our approach enables the reuse of suitable, already utilized explanation strategies, and can lead to an increase in the trustworthiness of AI systems in general.

The steps carried out to integrate this process within the CBR cycle are illustrated in Fig. 1. They are described as follows:

- Using the research literature and an analysis of use cases, we have created an ontology, iSeeOnto, which defines relevant concepts needed to describe an explanation experience. We describe the ontology in more detail in Section 3.1.
- We leverage iSeeOnto to gather and describe relevant features of best practice explanation experiences from different real-world use cases. The case base then contains complete records of historical explanation experiences defined using iSeeOnto concepts — the context description D, the solution S, and the outcomes O of evaluating the explanation experiences with end-users. We describe novel aspects of the case structure in Section 3.2. In particular, the case solution is a combination of explanation algorithms, formalized as an explanation strategy and represented as a Behavior Tree (BT) [38].
- This knowledge can then be accessed via similarity-based retrieval. A user's explanation need is formulated as a query formed only of the description component, D_q, according to the concepts defined in iSeeOnto. We use an implementation of CBR, CloodCBR [39], to compute the similarity between query context descriptions and the context descriptions of the cases in the case base, and we retrieve matching cases. This retrieval process is explained in Section 3.3.

- The explanation experiences that are retrieved may not be suitable or applicable to the query. For example, the explanation strategy that was retrieved might use a model-specific technique that is not transferable to a different AI model in the query. Here, the reuse process adapts the retrieved explanation solutions to the requirements defined in the query. Reuse and revise operators are used to address any mismatched requirements. We describe the reuse process in Section 3.4.
- The last step of iSee's CBR cycle is the revision of the explanation experience, allowing design-users to manually modify the retrieved strategy to meet any additional requirements. Following the revision process, the explanation strategy can be deployed in the evaluation cockpit. Then, end-users can receive explanations and provide an evaluation through conversational interactions. This process is explained in Section 3.5.

3.1. ISeeOnto: Formalization of the explanation experiences

The formalization of the explanation experiences, $\langle D, S, O \rangle$, is based on a semantic representation defined in the ontology *iSeeOnto*.[2] Fig. 2 illustrates its top-level concepts, and Table 1 summarizes the attributes used for case representation in iSee, organized through the iSeeOnto concept Case Feature. The description D includes the concepts related to the end-user, the AI model to explain, and the explainability requirements defined by the design-user. Fig. A.9 shows the user interface in the iSee platform that is used by the design-user to specify this information following the conceptualization defined in iSeeOnto. The solution S is formalized as a Behavior Tree (BT), a mathematical model whose execution is carried out in a modular fashion. O is the outcome of applying the experience according to the user evaluations. The concepts that describe the AI model are detailed in iSeeOnto. Specifically, it includes the AI task to solve, the AI method that the AI model employs, information about the dataset on which the AI model was trained (such as the type of data), and information about the AI model performance [16].

The *explainability requirements* are also defined in iSeeOnto. These concepts define the explainers (explanation methods like LIME, Anchors, Integrated Gradients, etc.) that the design-users need or want to be included in the explanation solution. Most explainability requirements concepts are well-known in the XAI literature: explanation scope, portability, and concurrentness. We also have other concepts such as the explainability technique required, the explainer implementation framework, its computational complexity, the explanation target, the data type that the explainer can process, the AI task, the AI Method that it can explain, and the type of presentation (output format) used to show the explanation.

Regarding the *end-user* concepts included in iSeeOnto, we have the following: the question that the user wants to answer when getting the explanation, the intent associated with the question, the question target, the user domain knowledge, the user's knowledge of AI, and the user's technical facilities. From this formalization, the iSee platform lets design-users define different *user personas* that represent prototypical users of the AI system with concrete explanation intents. From the literature, we know that user personas are archetypes of users that are designed to help satisfy users' needs and requirements [40]. Therefore, in the context of the iSee platform, it is crucial to allow design-users to define use case personas by mapping the mental models of users to specific AI contexts. Fig. A.10 shows the corresponding user interface.

The solution of an explanation experience is formalized as a BT; this is where the explanation strategy execution is defined. The solution has one explainer to execute at least, but it also can have multiple explainers which are going to be executed following the workflow defined by the BT. To do that, the design-user can create a BT from scratch or modify one already created (for example, when a solution is suggested after the retrieval step). An example of a BT for iSee is in Fig. A.12; it will be described further in the next section.

Finally, the outcome O captures the end-user's evaluation of the explanation experience. It focuses on assessing the extent to which the explanation satisfies the requirements defined in the description D.

Table 1
Case representation attributes of iSee's CBR system. A combination of characteristics of the AI model, target explanation, and end-user are utilized to describe a given explanation experience.

	Attribute	Example values
AI model	AI task	Classification, Regression
	AI method	Neural network, Random forest, SVC
	Dataset type	Tabular, Text, Images
Explanation	Portability	Model-agnostic, Model-specific
	Scope	Local, Global
	Target	Model, Instance
	Presentation	Image, Chart, Table
	Concurrentness	Ante-hoc, Post-hoc
User	Intent	Performance, Transparency, Trust
	Technical facilities	Touch screen, Mouse, Speaker
	AI knowledge level	Low, Medium, High
	Domain knowledge level	
	User questions	–
Solution	Explanation strategy	–

3.2. Definition of explanation strategies through Behavior Trees

The iSee methodology requires a rich structure to represent the workflows within the explanation strategies — one that supports reuse across different scenarios. In iSee, we formalize explanation strategies using BTs supported by knowledge derived from iSeeOnto.

Although BTs were first created as a tool to enable modular development of AI in computer games [41], in the last decade they have received increasing attention in other communities [38,42]. In BTs, the workflow logic is not dispersed across many states, as it would be in state machines, but organized in a hierarchical tree structure, with the actions as leaves. This has a significant effect on modularity, which in turn simplifies both synthesis and analysis by humans. BTs also provide scalability, reusability, and comprehensibility as they make it easy to express needs in terms of goals to achieve. These advantages are needed in games AI design and other fields.

BTs are hierarchical structures consisting of labeled nodes designed to be executed by an agent. The traversal of a BT starts from the root node and progresses through the tree by executing actions in a predetermined manner until a terminal state is reached, indicating success or failure. Within BTs, leaf nodes are referred to as *execution nodes*, representing specific executable behaviors. Each leaf node performs a particular action, ranging from simple checks to more complex tasks, and produces an output status, such as *success*, *failure*, or *running*. These leaf nodes establish the connection between the BT and the underlying code of the application in which the agent operates. In the context of iSee, the leaf nodes represent explainers, providing explanations within the system.

An example of a BT for iSee is in Fig. A.12. This BT illustrates a real use case that explains a neural network that aims to identify defective parking sensors. For this concrete scenario, we can identify two potential users: the AI engineer and the manufacturing inspector. The AI engineer focuses on improving the performance of the current predictive system and identifying potential biases and erroneous behaviors. The manufacturing inspector's role is to verify the quality of the sensors that have been potentially misclassified by the system. While both users may have different intents, they can benefit from the explanation methods included in the complete explanation strategy.

[2] Available at: https://github.com/isee4xai/iSeeOnto. Documentation available at https://isee4xai.github.io/iSeeOnto/docs/explanationexperience-en.html.

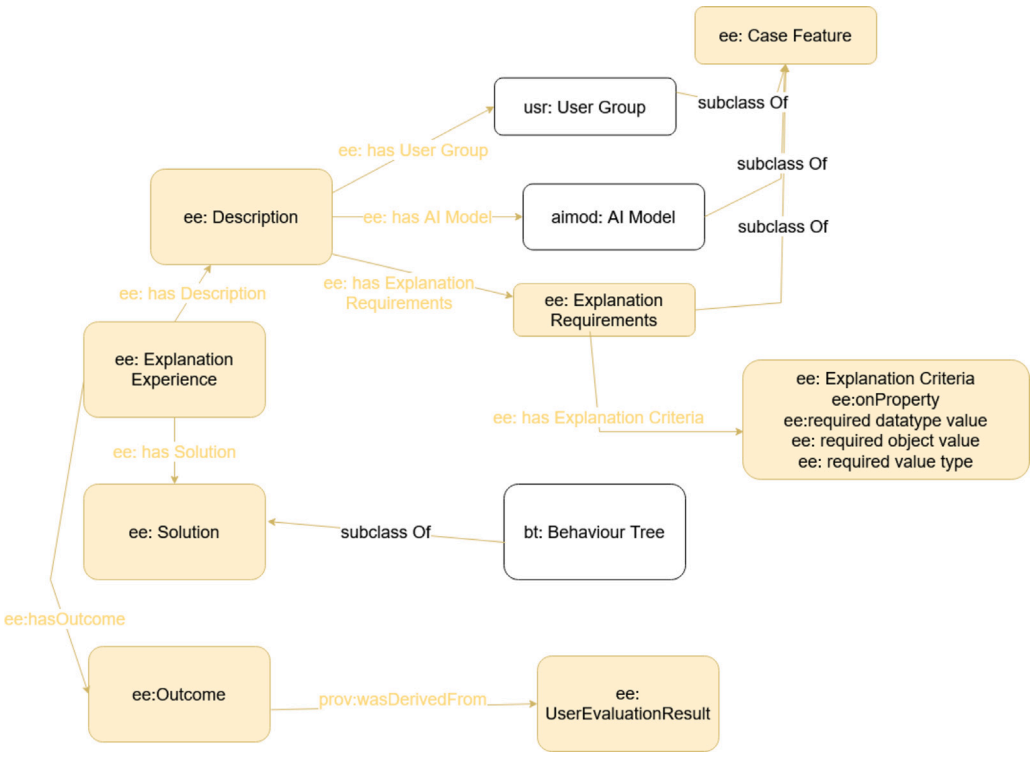


Fig. 2. General overview of iSeeOnto. An explanation experience is formalized as a tuple $\langle D, S, O \rangle$, where D is the description of the end-user, the AI model to explain, and the explainability requirements; S is the solution represented as a BT; and O is the outcome of applying the experience according to the user evaluations.

The first sub-tree in Fig. A.12 provides an overview of the performance of the model. This section of the Explanation Experience is mostly oriented to the intent of the AI engineer. This user may ask questions such as: "What is the recall of the system for defective sensors?" or "What is the percentage of sensors that tend to be mislabeled?"'. The execution flow of the experience provides a summary of the performance of the model by showing the precision, recall, and f1-score of each class. If the user would like a complementary explanation afterward, the confusion matrix for the test data is provided.

The second sub-tree in Fig. A.12 contains local explanation methods based on feature attribution. In other words, these explanation methods aim to identify the pixel regions that have a higher impact on the classification of a particular instance. This kind of explanation method may fulfill multiple intents at once for both of the defined end-user profiles. For example, manufacturing inspectors may ask questions to help them assess the quality of a sensor, such as "What region of the sensor has been identified by the system as defective?". On the other hand, the AI engineer may obtain valuable information about the behavior of the model for troublesome instances, asking questions such as "What parts of the image have a greater impact in the classification of this sensor?" or "What image features tend to be more important according to the system?". The explanation methods are presented in the following order until the user is satisfied with the explanations or all the explanation methods are executed: Grad-CAM, Integrated Gradients, LIME, and Anchors.

The top-right sub-tree in Fig. A.12 provides instance-based explanations for a particular prediction through the K-Nearest Neighbors and In-sample Counterfactuals explanation methods. Instance-based explanations tend to be more useful for end-users who may ask questions such as "What images would get a similar/different outcome to this one?" or "What does an image of a certain class look like?". However, in this scenario, the explanations are not targeted at end-users, but rather at expert users such as AI engineers and manufacturing inspectors. Nevertheless, the explanations in the top-right of the figure might still be useful. For example, when a sensor presents a new kind of defect that considerably differs from the patterns learned by the predictive model, the neighbors shown by the nearest neighbors explanation method would be ones with low similarities to the new defect. If this situation happens repeatedly, it may be an indicator for the AI engineer that the model should be retrained on a broader dataset, including instances with this new type of defect.

While this example has been manually designed for the sensor anomaly detection use case, it can be reused with different image classification models. Although the covered explanation methods are mostly model-agnostic, techniques such as Grad-CAM and Integrated Gradients can only be applied to convolutional neural networks, so the strategy might require some adaptation if a different AI model is used. Nevertheless, the flow may be applied to AI models for domains other than the manufacturing industry and quality control, such as healthcare, security, etc., where AI models are also commonly trained to perform image classification. This way, BTs demonstrate their suitability to support the reuse of explanation experiences across different domains.

3.3. Retrieval of explanation experiences

The process of explanation experience retrieval starts with a query, for which we only know its description D_q. Our ultimate goal is to find the solution S for the query. Within the retrieval step, we find some solution candidates, which are the solutions of the most similar cases to our query. To find the most similar cases, we compare our query description D_q with the descriptions of the cases in the case base. To

make this comparison, we analyze all attribute values in D_q against the case descriptions in our case base.

Attributes in D are formalized by the iSee Ontology. An attribute value can take three formats: a class, an individual or a set of classes/individuals. The local similarity metric for each attribute is selected based on the attribute format. Specifically, we use the *Exact Match* metric to compare individual instances; *Query Intersection* to compare sets of classes/individuals, and *Wu & Palmer* [43] to compare classes, which computes similarities taking the ontology hierarchy into account. The global similarity between two cases is calculated as the aggregation of all local similarities. We implemented the retrieval process with the CloodCBR [39] framework. More details about the retrieval process can be found in [43,44].

Figs. A.9 and A.10 illustrate how this process is performed in the iSee platform. To generate a new use case to explain, the design-user must add all the information related to it (i.e. the AI model to explain, explanation requirements, and end users' profiles), which is the knowledge we modeled with iSeeOnto as explained in Section 3.1.

Thereafter, design-users receive different explanation strategies as solutions (for a specific end-user and intent) in accordance with the information collected in the previous step (Fig. A.11). These solutions are explanation strategies that were found to be satisfactory and hence were applied in previous similar use cases.

3.4. Reusing Behavior Trees

In the reuse step, we start from the list of solution candidates associated with the most similar cases to our query (computed as explained in the previous section). The design-user can choose one of those solutions (a BT) as it is. These solutions may already be usable for the query since the case attributes are considered when calculating the similarities. Some of those attributes are user requirements. Therefore, the solutions are already user-specific at that point. However, design-users may want to change a part of the BT that is not applicable to the requirements or could be adapted to better satisfy their requirements. Therefore, the reuse step is the procedure where the design-user can change those parts in the BT that need to be modified, either because that part is not suitable for the use case or because the user would rather carry out different processes and use different explainers.

In the iSee platform, we have developed a tool, the Explanation Experiences Editor (iSeeE3) [45], to create new BTs from scratch or to edit them manually. Regarding the reuse step, iSeeE3 allows design-users to perform two main tasks: (1) to find similar and applicable explainers to the ones that appear in the BT solution that the design-user wants to modify, and (2) to find similar BTs to a subpart of that BT solution. For both procedures, the design-user can specify explicit requirements through a form shown in the BT editor. With this functionality, design-users can select which explainer properties (from the ontology) they do wish to have in their BT solution. For example, they can ask to include explainers that show explanations in text format. For this reason, the solution they build is user-specific, as they can choose the substitutions they want and the explainer properties they need.

For the first task, the design-user can access a set of applicable and similar explainers to the ones that the current BT has in its leaves. The choice is obtained from a list of explainers that we keep in the iSee *Explanation Library*, a unifying API that integrates (at the time of the writing) 67 explainers from different XAI libraries [24]. To get the applicable explainers according to the requirements described in the query, we perform a filtering step, where we check the properties that affect the proper working of a specific explainer, given a specific use case and context. These properties are: the dataset type that the AI model explains and the explainer can process, the AI method type, the AI task type that the explainer can explain, and the explainer implementation framework. At this point, if the user has included explainer requirements, we also perform a second filtering step, where only the explainers that fulfill the user's design requirements are retrieved. After that, among the explainers that are applicable, we find the most similar ones to the explainer that the design-user wants to replace. The similarity measure we use for this comes from a study that we conducted in which we compared different semantic similarity measures and analyzed the effect of using ontology knowledge. This similarity metric is called *depth* and considers the hierarchy of the explainer properties in the ontology to get the similarity values. Finally, we show the design-user the list of applicable and similar explainers. Users will pick the recommendation that they consider the best for their user case, i.e. transforming the solution by replacing explainers in their BT to give a user-specific solution. Additionally, the design-users can view a natural language explanation that presents the explainer properties that the current explainer and a similar explainer have in common.

For the second task, the design-user selects the root in the BT, and the iSeeE3 retrieves the most similar BTs to the tree query. We look for similar BTs among all the solutions saved in our case base. Here, we apply the Levenshtein edit distance [46], which computes the cost of transforming one BT into another one considering the insertions, deletions, and node substitutions that would be needed. The Levenshtein edit distance is a sequence edit distance originally for strings [47]. In iSee we use an adapted version of this metric for graphs. This version comes from an edit distance library.[3] We represent each BT with a list of nodes and an adjacency list that depicts the children for each node within the BT. The adapted Levenshtein edit distance compares these sequences to get a score that symbolize the differences between two BTs [46]. The cost of transforming one BT into another is used as a similarity measure since we pick the BT with the lowest distance to our subtree. Once the similarity functionality is executed, we offer k solutions (similar BTs) to the design-users so they can choose the option that they want. Again, users will pick the solution that they think best suits their use case, so they are again getting a user-specific solution.

Moreover, both reuse processes can also be done automatically through two different buttons on the editor. On the one hand, when there are explainers within the BT that cannot be applied to the design-user's particular use case, the user can click on the "Fix inapplicable explainers" button, and all those inapplicable explainers will be replaced by their most similar explainers that are applicable. On the other hand, the "Adapt subtree to be applicable" button can be used whether the explainers in the BT are applicable or not. If the user chooses this button, the subtree is replaced by the most similar BT from the case base whose explainers are applicable to the use case.

The reuse process is exemplified in Fig. A.13. The panel on the right side shows candidate alternative methods for a given explainer.

Finally, the BT solution, modified by the design-user, is deployed to be evaluated by end-users through a chatbot, as explained next.

3.5. Deployment of the explanation experiences

When the design-user finishes constructing a solution, the end-user can access a chatbot in the iSee platform to try out the explanation obtained with the solution and evaluate it. The chatbot follows the execution workflow defined by the BT. The end-users will answer questions about themselves (their profile) and their intent, and they will receive the explanations, as shown in (Fig. A.14).

The conversation is conducted by executing the BT from left to right with the following steps: begin the interaction by greeting the user and obtaining permission to continue; identify the persona (end-user profile) based on knowledge levels; determine the explanation

[3] The library code for the Levenshtein edit distance adapted for graphs that we used is here: https://gitlab.ub.uni-bielefeld.de/bpaassen/python-edit-distances/-/blob/master/edist/sed.pyx?ref_type=heads.

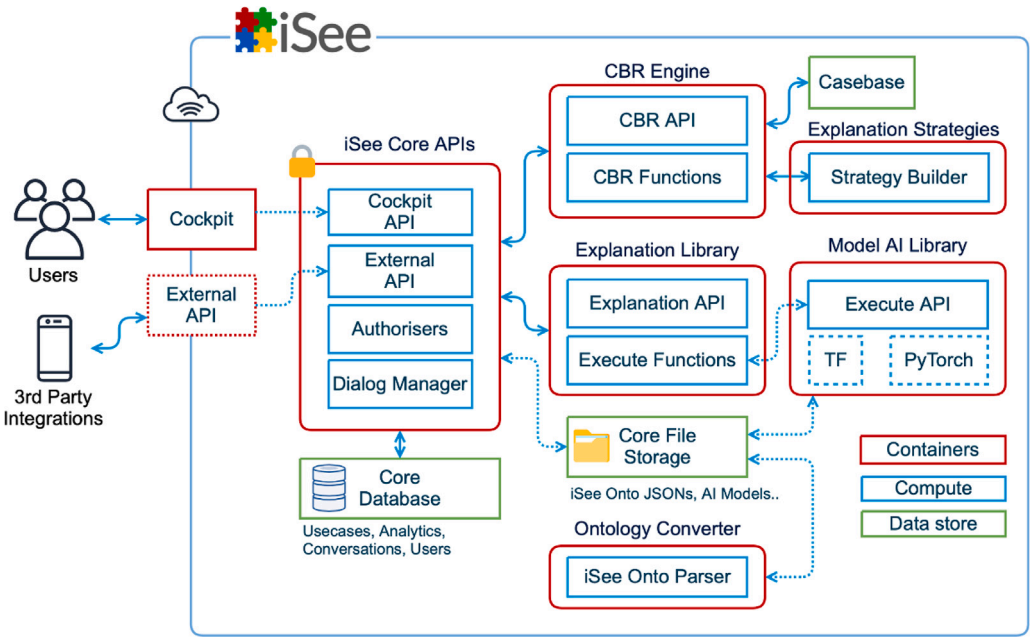


Fig. 3. Overview of the iSee platform.

target, i.e. the data instance and its AI system outcome; ascertain the user's explanation need by asking questions, and provide explanations to answer those questions by running the appropriate explainers of the explanation strategy; this is repeated until the user has no more questions or the XAI system is unable to answer any more questions; assess the experience using the evaluation questionnaire; and finish the explanation experience conversation.

3.6. Outcome evaluation

The feedback that was collect from the end-users during the conversational interactions forms the outcome of the use case. The end-users respond to questionnaires designed by their design-user. The questionnaires can be designed to understand one or more evaluation dimensions such as user satisfaction, explanation quality, or trust. Fig. A.15 shows an example of how the design-users can define an evaluation questionnaire to assess an explanation strategy for a specific intent and end user.

The analytics module provides an interface (Fig. A.16) to view a summary of end-user feedback grouped by evaluation dimensions. Design-users may use this information to optimize the explanation strategy by understanding user perceptions of the current solution and identifying unmet explanation needs. After several iterations, when the end-user experience achieves the desired level of user satisfaction, the design-user can finalize the explanation strategy for deployment. At this stage, the iSee platform may retain an anonymized version of the use case in the case base for future reuse.

4. The iSee platform

The architecture of the iSee platform is illustrated in Fig. 3. It consists of several tools that implement the CBR methodology presented in the previous section. Its architecture is modular, based on several modules that provide different APIs and that enable the expansion of the platform. For example, the *Ontology Converter* allows the system to dynamically access the knowledge contained in iSeeOnto, such as the case representation and similarity knowledge. The AI models to be explained and provided by the platform users are stored in the *Model AI Library*, which is accessed by the Explanation Library, a unifying API that integrates explanation methods from different XAI libraries [24]. Design-users can submit a new explainer for inclusion in the library if that explainer is not present. A complete repository of explainers is stored on GitHub.[4] For an explainer to be added to this repository, it must be described using the ontology (as shown in Fig. A.17).

The whole iSee platform is based on a micro-services architecture and each component of the platform is executed in its own isolated environment (container). A multi-tier network architecture has been implemented to enhance the security of the whole solution by limiting the network access to and from the different services:

- **A front-end network**: This is the only network that is connected to the Internet and is used to host a reverse proxy service (nginx) to direct incoming traffic to the appropriate applications.
- **An application network**: All applications containers are connected to this. No direct access from the Internet is allowed from this network and applications are never exposed directly to the Internet.
- **A database network**: This contains all database services used by the iSee Platform. This network is only visible from the application's network.

The web traffic to the different applications is encrypted by SSL/TLS protocols; the reverse proxy service is in charge of providing this level of security on behalf of all applications. Note that the security certificates are periodically renewed (every 90 days) by an automated process. The whole platform is also hosted in a secure environment, behind a firewall that prevents common threats and attacks.

The iSee platform implements authentication/authorization through JSON Web Tokens (JWTs), which is an industry-standard to securely send data between two parties — in our case our servers and the client's web browser. Use cases are only accessible to users of the same

[4] https://github.com/isee4xai/iSeeExplainerLibrary.

Fig. 4. Example image examining the patient's forearm. An oblique fracture is present in the ulna — meaning that the bone is separated, but the two parts are not displaced.

organization and the case base is completely anonymized — meaning that, when the cases are retained, no information that identifies the source of the use case is saved.

5. Evaluation and use cases

We have conducted experiments that aim at evaluating whether iSee helps the design-user to assess the outcome of an explanation strategy, and how effective the tools presented in this paper are in helping the task of designing a suitable explanation strategy and understanding how far the strategy meets the explanation requirements of users. The research questions we look at are:

- **Usability**: does the iSee platform help in assessing the end-user's perception of a given AI system?
- **Meaningfulness**: Do the end users perceive the explanation conversational experience fit-for-purpose?

We have carried out evaluation experiments with a diverse community of 20 engineers, managers, and desk agents. Although they were not experts in the domain we used in the experiment, all of them had information technology experience and a technical background. The participants were asked to use iSee tools in one real-world use case within medical radiography fracture detection. To overcome ethical issues, we informed the users about the nature of the experiment, its goals, the tasks they had to carry out, and how the results were going to be used. No personal data was collected. Therefore, our experiment ensures the participants' autonomy, privacy, and confidentiality, and provides them with trustworthy information.

The study was organized according to the following protocol:

1. We provided a brief overview of the iSee project and an online tutorial of the tool on one of two real world examples.
2. We shared a use case sheet detailing the AI model objectives, the explanation goals and examples of data instances and expected

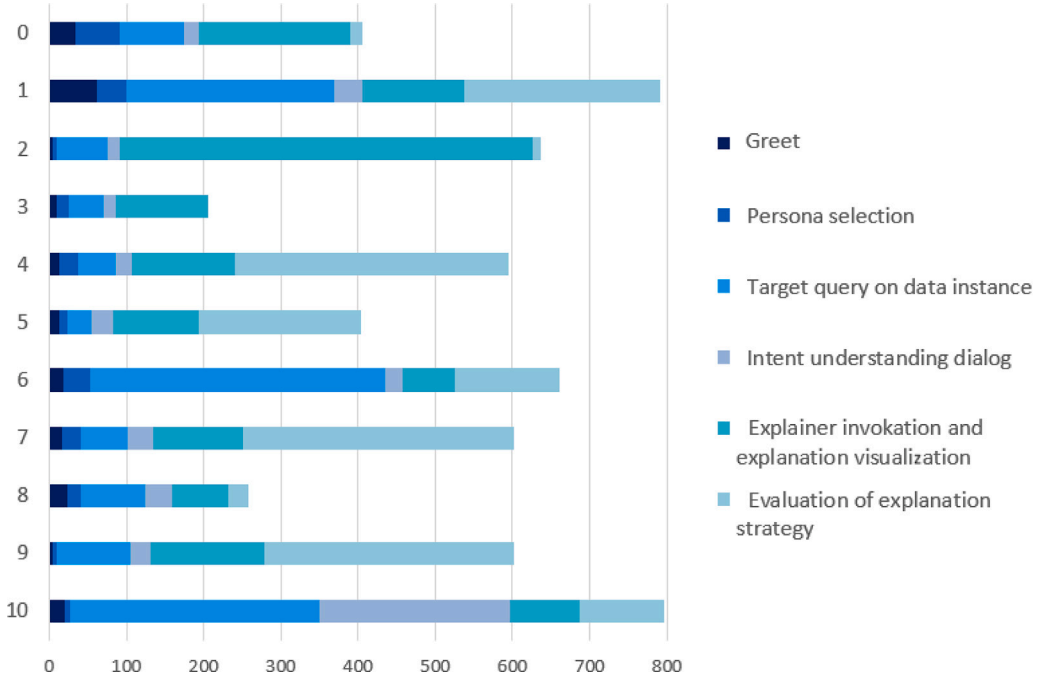


Fig. 5. Time-in-secs breakdown per individual user and step. The X-axis represents the seconds spent by each participant in each step. The Y-axis represents participants' ids who participated as clinician.

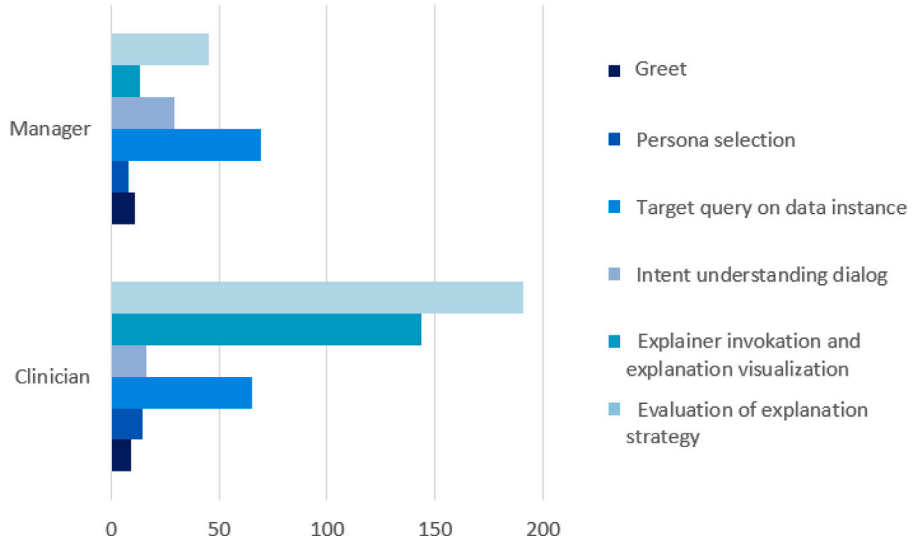


Fig. 6. Time-in-secs breakdown per persona. The X-axis represents the average time in seconds spent, while the Y-axis represents the step in which the participant (as clinician or manager persona) stay during this duration.

results. The participants were given a list of tasks to perform and invite link to access iSee.

3. We then asked the participants to follow the invite link and interact with the iSee chatbot. The completion of the entire session was key for us to record the full user experience and be able to analyze accurately the link between the user profiles, the visualization of produced explanation in the chatbot, and the user perception.
4. Finally a qualitative assessment, using a mixed method study, was performed to evaluate the interactive experience of end-users.

5.1. Summary of the use case

The AI task is the analysis of medical documents (radiographic images) to support a medical expert (clinician) or a patient (self-analysis). The use case aims at simulating the real world: the AI model receives radiographic images from medical imaging equipment (in this case, an X-ray machine). The goal of the AI system is to detect whether or not a fracture is present in the X-ray image. Many different types of fracture may be present in the data (i.e. different fracture types, different stages in the healing/treatment process, etc.), but the AI needs

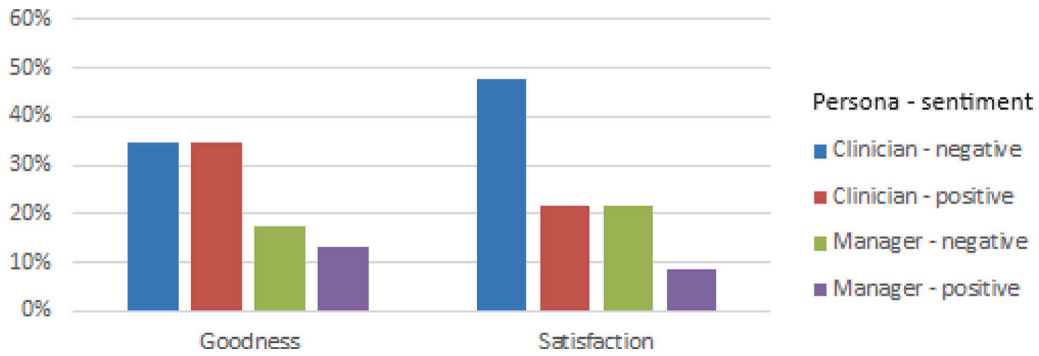


Fig. 7. Assessing whether the produced explanation meets persona-wise evaluation goals. Each column represents the percentage of interactions carried out with each persona together with the sentiment produced for Satisfaction and Goodness. For example, the blue and red columns represent the interactions done when the participants acted as clinicians. The blue ones are the percentage of interactions perceived as negative, while the red ones are the percentage of interactions perceived as positive.

only to predict whether a fracture is present or not in the image. Fig. 4 illustrates this.

Two personas were designed:

- *Clinician*, where we set the goal of the explanation to be: supporting in assessing the result and deciding the next task (accountability); and
- *Manager*, where we set the goal of the explanation to be: helping understand the reason why the AI system produces its result and how accurate the system is (transparency and performance) as shown in Fig. A.10.

At the evaluation stage, the end-users were told to choose a persona and a data instance, select the question that reflects their goal and walk through the results exposed by the chatbot, and finally answer the evaluation questions posed by iSee.

There are three explanation methods used in the strategies proposed in this use case:

- A feature importance-based explanation, which overlays the query image with a heat map to indicate which pixels contributed to the AI model's classification (Integrated Gradients).
- An example-based explanation, which shows similar images to the query to explain why an image was considered to have a fracture or not (Nearest Neighbors).
- A performance-based explanation, which describes the performance of the AI model as a whole, using suitable metrics.

In the evaluation, iSee recorded the following: the persona that the user choosed, the explanation the user viewed, the order in which they were viewed in the interaction sequence, the evaluation questions and responses, and the time recording of the interaction steps. We then used sentiment analysis and a specific user questionnaire as a method for obtaining qualitative insights into user experience and the users' thoughts as they use the chatbot with a specific query and goal. The experiments generated 3 artifacts: (1) time recording of user experience, (2) a User Experience Questionnaire, (3) the records of each interactive session. The session records include, for each user, the intent question selected by the user, the explainers used, and at which position in the sequential script of the conversation (whether it was the first explainer used in the strategy or the second), and finally the evaluation question by iSee and answer by user for each explanation. We applied the last two artifacts to the results with the aim of evaluating the perception by the users (positive or negative) at each step of the interaction. We use two measures – Goodness and Satisfaction –, and we highlight potential relationships between the position of explanations in the interactions and the users' perception of them.

The measures designed in iSee for this use case come from the library of evaluation metrics in [32]: Goodness as the quality evaluation metric, which is more objective, and Satisfaction, which relates to a more subjective opinion from users.

5.2. Highlights & discussion

The following highlights were observed:

- Overall, the volume of Clinician sessions and Manager sessions are equally balanced. Clinicians asked more about which instance gets a similar outcome and what features the AI model considers, which led iSee to use IntegratedGradients slightly more frequently than NearestNeighbors, as shown in Fig. A.16.
- We used the time recordings produced by iSee to analyze the progression through the steps per user profile. Fig. 5 presents the time recordings broken down per individual user and step.
  Despite the freedom of the study, all users having chosen the same persona converged to a similar progression. Fig. 6 illustrates the findings where domain-aware users (clinicians in this case) focus more on evaluating the explanations whereas novice users (managers in this case) focus more on analyzing the result of the AI model.
  In the majority of the cases, we observe that users spend half of the total time checking explanation and responding to evaluation questions.
- We analyzed whether there is any relationship between domain knowledge and the evaluation of explanations by end users. The Clinician persona is set up with three evaluation questions under Satisfaction and three questions under Goodness, while the Manager persona has one question for each measure. We thus expect to see a one-third ratio between the results for the two personas. Fig. 7 shows that explanations do, to a degree, support users in assessing whether the AI model is good or not (Goodness). Moreover, perceptions are split in a balanced proportion between positive and negative, and this holds true for both personas. Domain and AI knowledge does not explicitly influence the results in this case. However, explanations are satisfactory to users in one-third of the cases when the persona is a Clinician, against a one-fourth ratio for the Manager persona. This suggests that domain knowledge is important to using the proposed explanation, especially when explanation strategies are visual (image-based) as in this example.
- We also analyzed the relationship between the position of the explanation in the interaction and the resulting user perception. This aims at helping to understand which explanation method provides sufficient details for an instance or needs to be complemented with another explanation. The evaluation is made for each question asked to the user since they reflect different aspects of suitability to user goals. In this paper we highlight results for two evaluation questions. Table 2 shows, for each explainer used, the position at which its result was exposed to the user and the total number of explanations viewed by the user within the same

interaction, and the perception derived from the user's evaluation result for the particular question selected. It is important to note that since the explanation strategy is personalized for each intent, the same explainer method could be applied with different parameters and at different position in the strategy, as shown in Table 2. On the top of the Table, the users were asked to judge the statement "The explanation of the AI model is sufficiently detailed". The overall perception is negative, although results for each explainer indicate that IntegratedGradients is more positively perceived (64% negative against 36% positive) than NearestNeighbors (86% negative against 14% positive). IntegratedGradients is clearly useful when it is used in the second position (at position 2, 100% positive perception) and rather neutral in other positions. This suggests that with this community of users, it would need to be combined with another explainer and is not seen as sufficient by itself. Conversely, NearestNeighbors is seen as not clearly bringing more information to the user compared to the previous explanation viewed (positive perception only appeared in up to 50% of the experiments). The Case-Based Reasoning system of iSee never recommended it in the first position (in the top table of Table 2 there is no position 1).

At the bottom section of Table 2, the iSee chatbot asked about the actionability provided by the explanation and whether it has helped know how to use the AI model. IntegratedGradients is positively perceived regarding this goal (60% cases with positive perception), with a higher result when it was viewed in the second position (at position 2, perception is 100% positive). The users' sentiment when being shown NearestNeighbors explanations is neutral no matter the position at which it was viewed by the user. This suggests that in this use case, it will work as a complement explanation. Specifically, the design-user would be advised to not use NearestNeighbors alone in the strategy during the adaptation.

- The study of the artifacts produced by iSee tools confirmed some important qualitative points. (1) The same explainer does not fit equally to all goals. (2) The perception of the explanation meaning and usefulness is dependent on the position of the explainer inside the sequence of explanations viewed by the interacting user. (3) An explanation not bringing utility when shown for the first time to the user, can be interpreted more positively when shown later in the sequence. (4) The participants reported the importance for them of guidance along the conversational flow so that they appreciate as sufficient the incremental information produced by the explanation compared to the previous one. (5) All participants found that Integrated Gradients based explanation was providing actionable and useful details when shown in second position, whereas the same explanation type was perceived as too "raw" format when proposed at the first position in the explanation strategy.
- The last analysis was performed using the User questionnaire and iSee records. In the User questionnaire we were interested in 4 psychometric dimensions related to establishing user experience and area of improvement. This was adapted from the XEQ scale for evaluating XAI Experience quality defined in [48]. The overall aim was to assess the user perception of utility and usability of the explanation tool:
  - **Learning**: Does the tool add value and new knowledge to the user?
  - **Utility**: Is the tool beneficial to the user task performance?
  - **Fulfillment**: Does the explanation achieve the intent desired overall?
  - **Engagement**: Does the tool contribute to provide what it is expected to (meet the intent)?

Each dimension measures the scores in 3 categories: positive, neutral, and negative. Overall the explanation visualized in the iSee chatbot and the conversation flow embedding the explanations have been scored positively for the 4 dimensions. While the limited number of participants does not allow us to generalize to statistical significance of this result, the results are very promising. Fig. 8 presents how the tool scored across the 4 dimensions measured by the user questionnaire.

Finally, The participants' feedback from the manual review allowed the design team to identify future work directions. On the one hand, the feedback helps the design users revise the explanation strategy, change the evaluation question, and combine global and local explainers. On the other hand, the format of the explanation is important: some end-users' feedback suggests that a visual explanation may be better understood when associated with textual, human-readable context information about the model or the data set. The overall experience would be improved if the explanations were preceded by a step guiding users on how to interpret the explanation or if they were able to select alternative explanations about the data instance they used. We have thus taken these remarks as a dimension for improvement to the tool.

Table 2
Assessing how the end-user perceives the usefulness of explanation: it can depend on when it is proposed during the interaction. We evaluate this for two different questions with their two different explanation strategies (top and bottom, respectively, in the Table).

Question 1: The explanation of the AI model sufficiently detailed			
Position of the explainer out of the total explanations viewed in the interaction	Total explanations	User perception: negative	User perception: positive
/Images/IntegratedGradients		**64%**	**36%**
1		**67%**	**33%**
	2	50%	50%
	3	50%	50%
	4	100%	0%
2		**0%**	**100%**
	2	0%	
3		**75%**	**25%**
	3	50%	50%
	4	100%	0%
/Images/NearestNeighbours		**86%**	**14%**
2		**80%**	**20%**
	2	100%	0%
	3	50%	50%
	4	100%	0%
4		**100%**	**0%**
	4	100%	0%
Grand total		**72%**	**28%**
Question 2: The explanation is actionable, that is, it helps me know how to use the AI model			
Position of the explainer out of the total explanations viewed in the interaction	Total explanations	User perception: negative	User perception: positive
/Images/IntegratedGradients		**40%**	**60%**
1		**40%**	**60%**
	2	0%	100%
	3	50%	50%
	4	50%	50%
2		**0%**	**100%**
	2	0%	100%
3		**50%**	**50%**
	3	50%	50%
	4	50%	50%
/Images/NearestNeighbours		**50%**	**50%**
2		**50%**	**50%**
	3	50%	50%
	4	50%	50%
4		**50%**	**50%**
	4	50%	50%
Grand total		**44%**	**56%**

5.3. Remarks

We can conclude from the evaluation that design-users can create the strategies, and get feedback from the end-users about the strategies

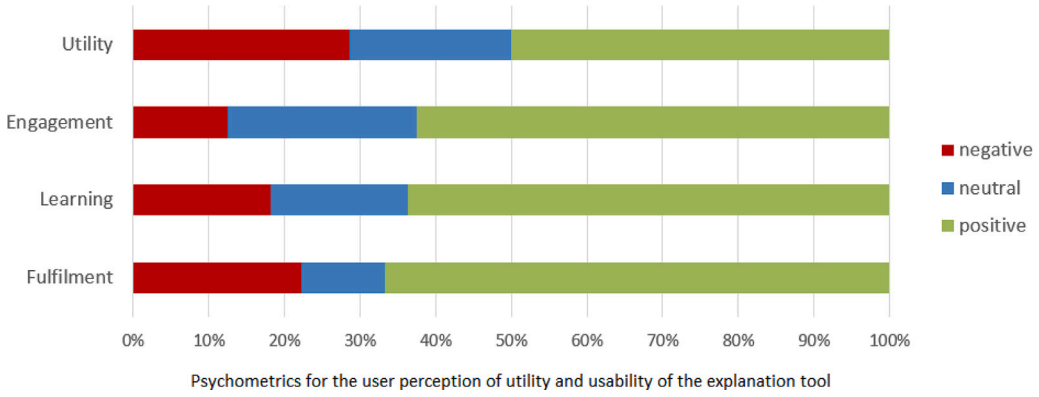


Fig. 8. Psychometrics for the user experience.

designed. The next stage of the evaluation experiments should include collaboration between design-users and end-users, enabling design-users to refine explanations based on end-user evaluations, so that the explanations better fit to the use case.

However, we want to note some ways to improve our evaluation in the future. Our use case is very domain-specific (medicine) but the users who participated are not experts in medicine. This might be the main reason why those users did not understand the explanations provided by the strategy. However, the goal of the experiments was to see whether the iSee tools provided the relevant information to assess user perceptions. According to the results, we can confirm that iSee is useful when designing explanation experiences and when evaluating them with real end-users. However, the next step should involve users drawn from the appropriate domain.

Moreover, another line of future work to improve the validation carried out here is to try different validations in different use cases and with a wider range of different end-user profiles. Our goal in the new experiments would be to enhance the generalizability and reliability of the results, together with the ethical issues identified in the following section that are related to the users' roles: confidentiality, auditability, or accountability. On the other hand, the issues related to security and accreditation should be evaluated by defining technical artifacts. The following section describes these issues.

6. Ethical issues

The scale of iSee's vision carries with it a number of ethical issues. In this section, we attempt to identify the interactions between different dimensions of these issues. Understanding these ethical issues helps us to de-risk usage of the iSee platform regarding privacy, auditability, and accountability.

The iSee platform is subject to a number of risks, similar to the union of those that would typically be faced by a large explainability framework and those of a multi-functional recommender system. One significant risk is a leak of confidentiality. This can happen at a strategic level (e.g. using intent understanding and intent similarity measures creates the risk of a data protection breach), at an individual use-case-level (e.g. the possibility that an explainer uses training data from another AI model), or even at a user-level (e.g. the explainer uses another individual's data to create an explanation within a single use-case). In some domains, avoiding the last of these might be a legal requirement (e.g. we should not use the explainer to retrieve non-anonymized data about another patient in a medical context). A confidentiality data breach is likely to be a risk in any scenario.

Based on the above, we have identified a number of ethical dimensions for the iSee system. While they are not all malevolent in nature, identifying the ethical dimensions will facilitate future threat modeling to safeguard the system.

- **Confidentiality:** Ensuring that iSee will have enough information to underpin its functionality while ensuring that no users are uniquely identifiable is a challenge, as explanations of models can explicitly or implicitly reveal individual data points [49].
- **Auditability:** Part of the platform's role is to support design-users to appropriately absorb and understand evaluation outcomes. This is a non-trivial problem, as a holistic evaluation of the explanation experience means understanding the interactions between complex and multi-faceted end-user goals [50].
- **Accountability:** Design-users are the nominated representatives of an organization and have a significant understanding of the organization. Therefore, they are accountable on the part of their organization. As part of this, ethical use of the iSee components will need to be controlled by a shared agreement that the user will have to sign and comply with.
- **Security:** Increasingly, AI frameworks are subject to threats from cyber attacks [51]. Since iSee supports the submission of new explainer algorithms to its library and of new cases to its case base, proactive policing of submissions, supported by reactive response to user feedback, will be a consideration in the deployment of iSee. Crowdsourcing feedback from users has been demonstrated in existing platforms (such as the Python Packaging Authority[5]) to be very effective, and is supported by a growing body of evidence in the literature [52]. Therefore, we plan to develop a user rating model alongside a suspicious behavior report form to quickly identify threats.
- **Accreditation:** Artifacts submitted to the platform should be acknowledged, and their impact documented for feedback to the original contributor. This will minimally require users of any explainer algorithm to provide appropriate accreditation for impact metric capture.

7. Conclusions

The iSee project was conceived as a way of helping AI developers who are looking for expertise to build explainability into their AI systems in line with regulations such as the right to obtain an explanation in the EU [53]. The necessity of explaining AI systems has increased in the last few years, with the appearance of black-box algorithms, like deep neural networks, in critical domains, such as health and cybersecurity. Many XAI methods have emerged in the literature, including LIME and Integrated Gradients. There is a huge

[5] The PyPa code of conduct highlights the community policing policy: https://github.com/pypa/.github/blob/main/CODE_OF_CONDUCT.md.

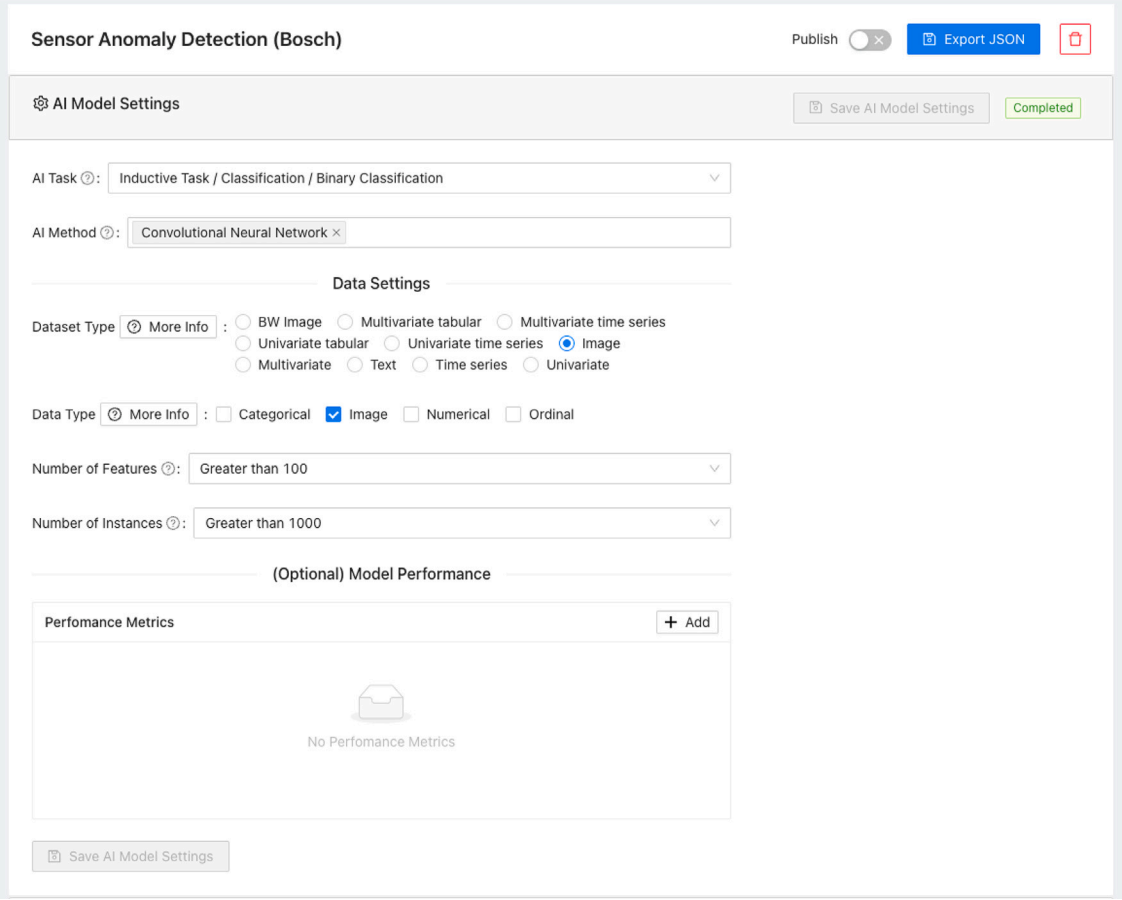


Fig. A.9. Interface to define the AI model for the sensor anomaly detection use case in the iSee platform.

number of explainers available, and picking out the best one is a difficult task, requiring design-users to balance many criteria.

The iSee platform tackles this problem by making use of the CBR methodology and relying on valuable user experience. In this paper we have described how iSee offers a set of different tools oriented to AI design-users, to help retrieve and reuse previous explanation experiences. The iSee platform aims to become a go-to platform for promoting trustworthy AI in both industry and research communities.

CRediT authorship contribution statement

Marta Caro-Martínez: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Conceptualization. **Juan A. Recio-García:** Writing – review & editing, Visualization, Validation, Methodology, Investigation. **Belén Díaz-Agudo:** Writing – review & editing, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Jesus M. Darias:** Writing – review & editing, Visualization, Validation, Software, Methodology, Investigation, Conceptualization. **Nirmalie Wiratunga:** Visualization, Validation, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization. **Kyle Martin:** Visualization, Validation, Methodology, Investigation, Conceptualization. **Anjana Wijekoon:** Writing – review & editing, Visualization, Validation, Software, Methodology, Investigation, Conceptualization. **Ikechukwu Nkisi-Orji:** Visualization, Validation, Software, Methodology, Investigation, Conceptualization. **David Corsar:** Visualization, Validation, Methodology, Investigation, Conceptualization. **Preeja Pradeep:** Visualization, Validation, Software, Methodology, Investigation, Conceptualization. **Derek Bridge:** Visualization, Validation, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization. **Anne Liret:** Writing – review & editing, Visualization, Validation, Supervision, Resources, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The link to GitHub is included in the manuscript.

Acknowledgments

This research is funded by the iSee project. iSee is an EU CHIST-ERA project which received funding for the UK from EPSRC under grant number EP/V061755/1, for Ireland from the Irish Research Council under grant number CHIST-ERA-2019-iSee (with support from Science Foundation Ireland under Grant number 12/RC/2289-P2 at Insight the SFI Research Centre for Data Analytics at UCC, which is co-funded under the European Regional Development Fund), for Spain from the MCIN/AEI and European Union "Next Generation EU/PRTR" under grant number PCI2020-120720-2, and for France under grant number ANR 21-CHR4-0004-01 (iSee CHIST-ERA Call 2019).

Appendix. The iSee platform interface

See Figs. A.9–A.17.

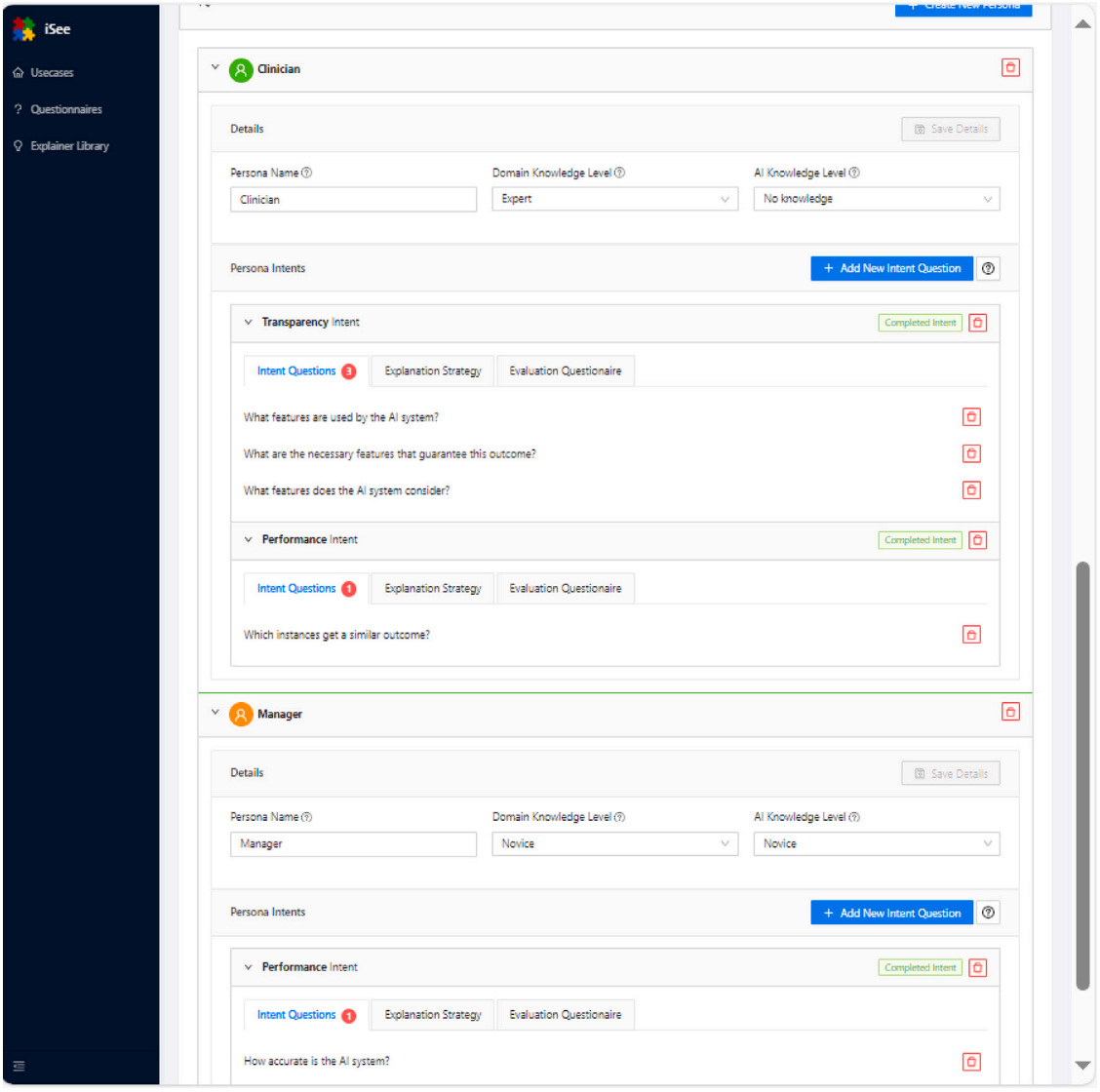


Fig. A.10. Persona and intent questions in the radiograph use case.

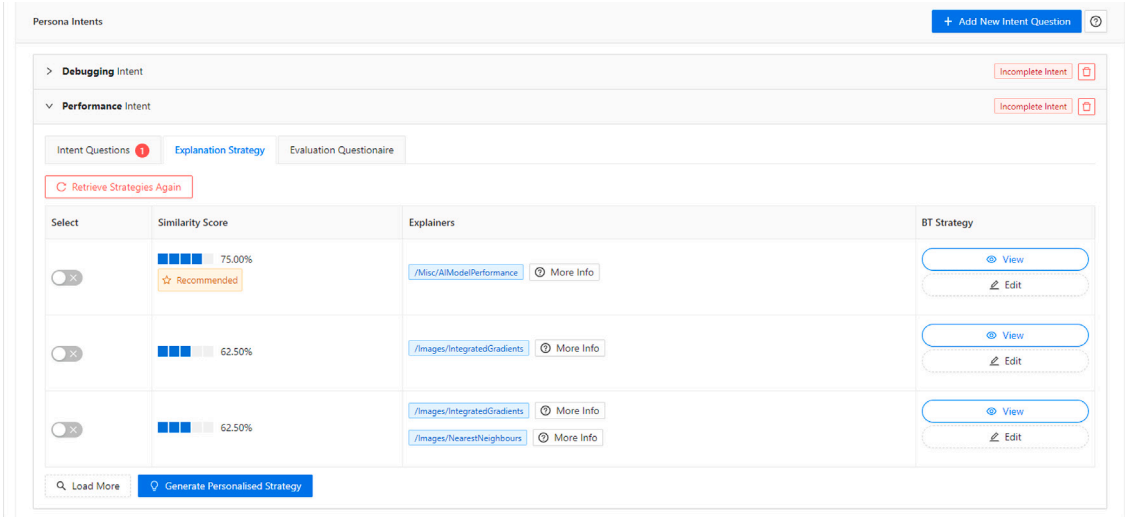


Fig. A.11. Retrieving possible explanation strategies for a specific user type and intent.

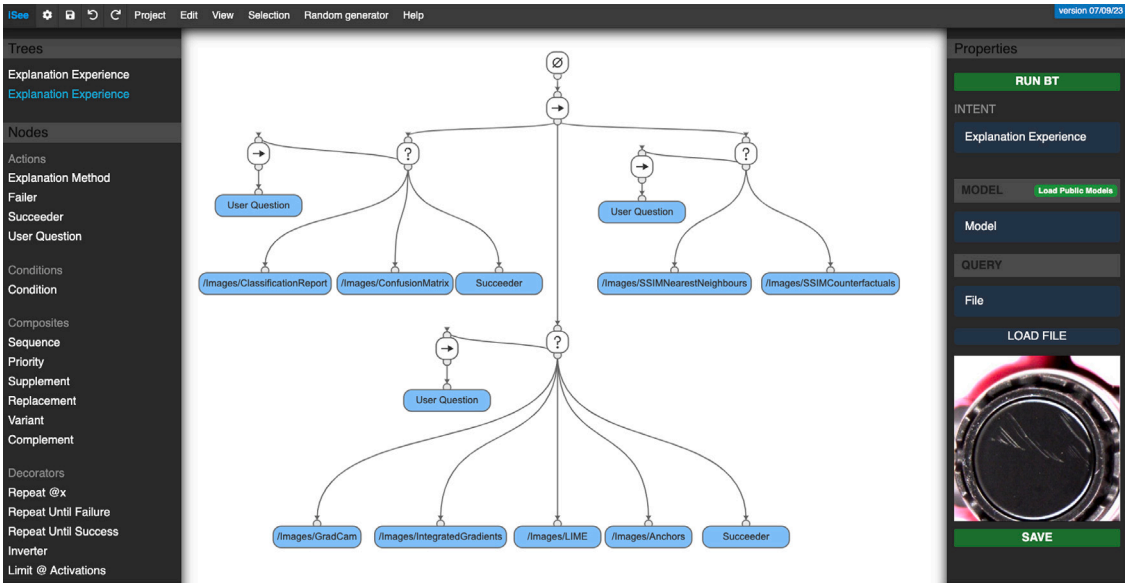


Fig. A.12. An example of a BT designed for iSee. Nodes labeled → are sequence nodes; those labeled ? are priority nodes. Apart from those, we also have other types of nodes defined to create a better explanation strategy process. Each explainer is associated with the user question that it satisfies.

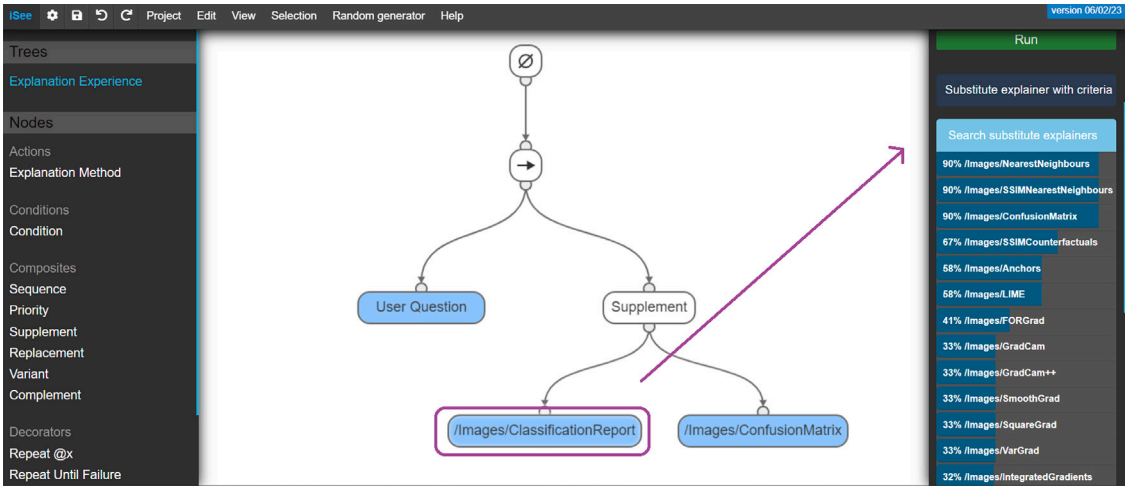


Fig. A.13. iSeeE3 suggesting to change an explainer in a solution considering different similar explainers.

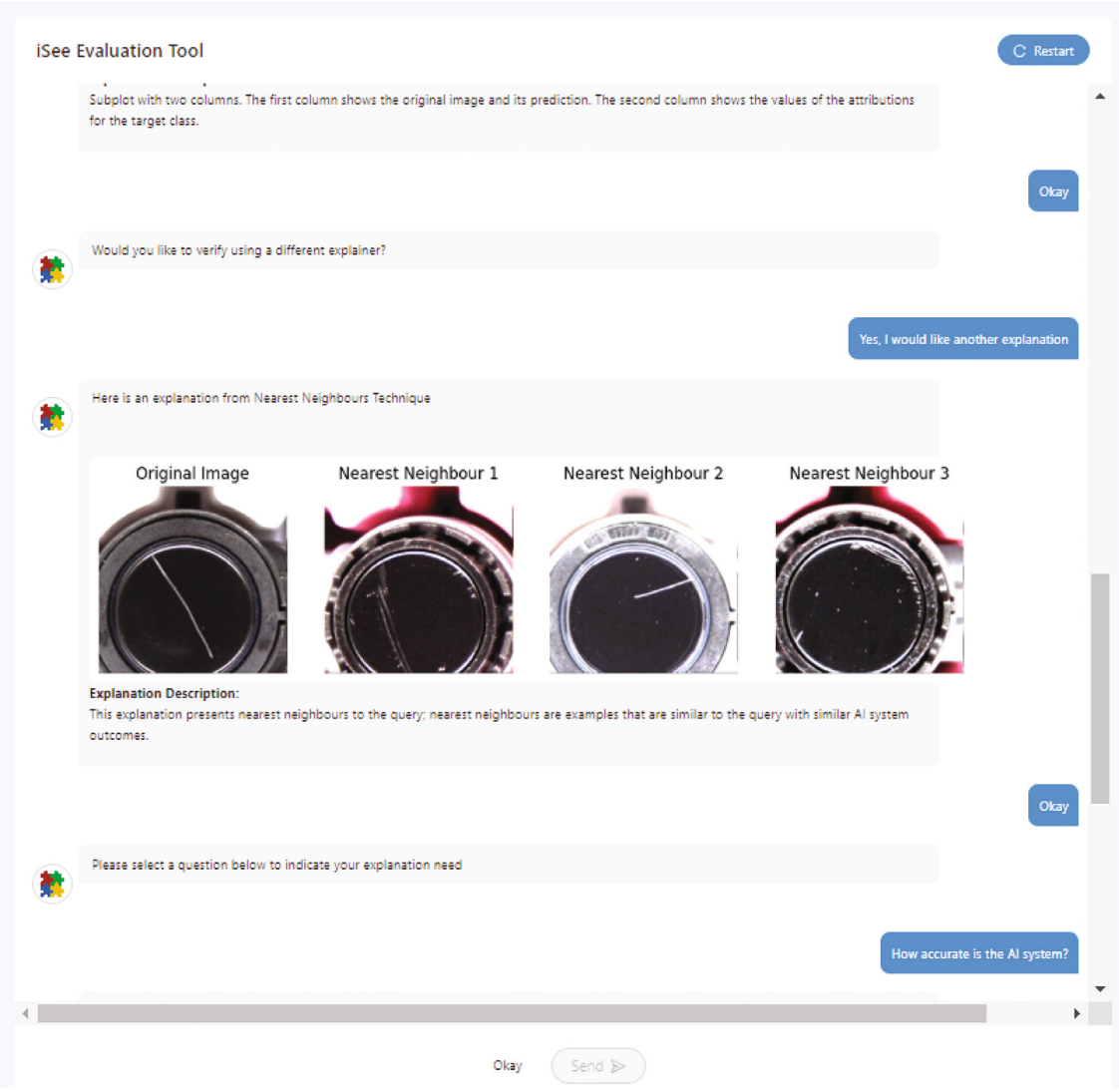


Fig. A.14. Chatbot to show the explanation solutions for the sensor anomaly detection use case to end users: example of a loan approval use case.

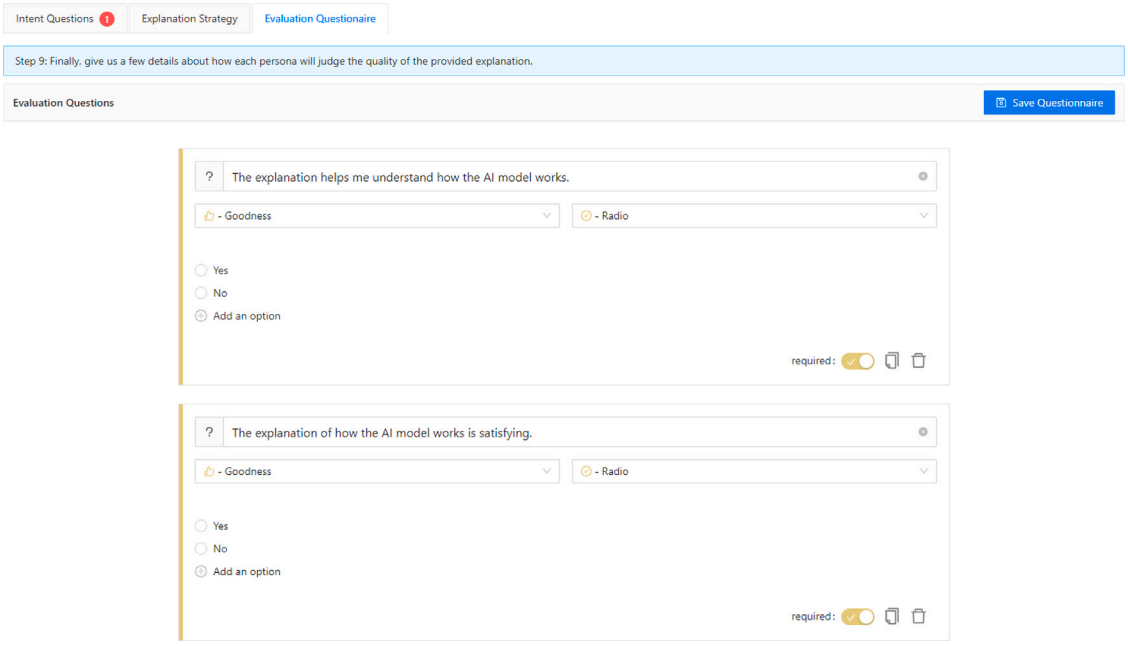


Fig. A.15. Screen to design an evaluation questionnaire for a sensor anomaly detection end-user and a specific intent.

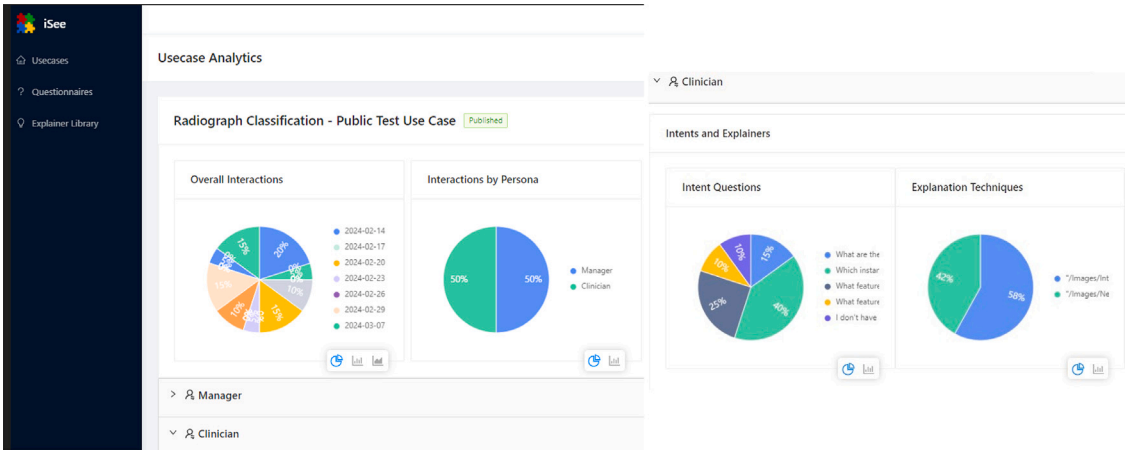


Fig. A.16. Fracture detection experiments: overall metrics of iSee.

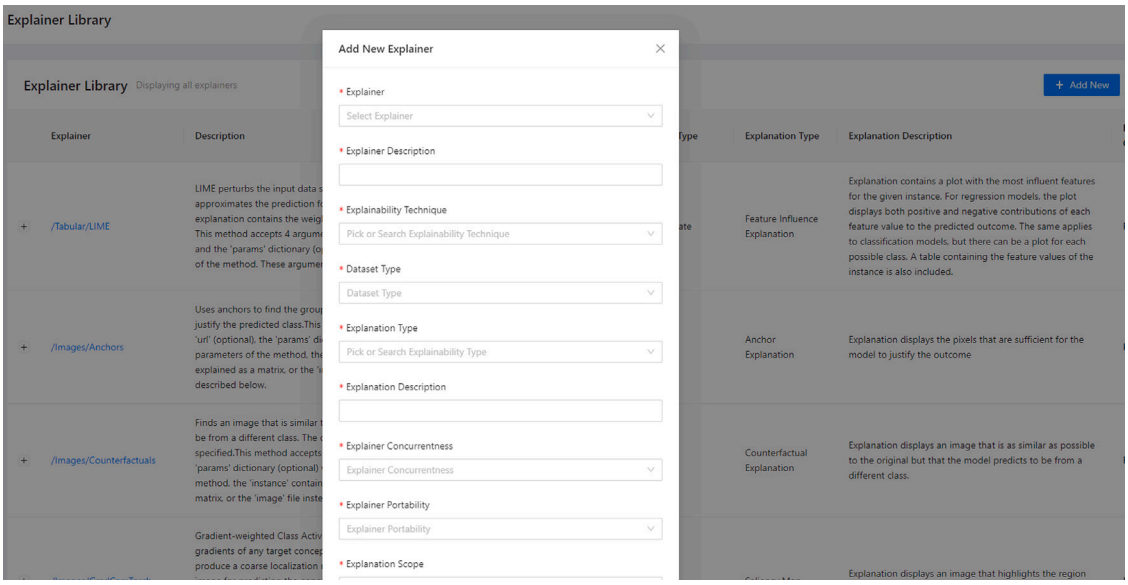


Fig. A.17. How to add explainer properties of a new explainer to our explainer library.

References

[1] K. Bera, N. Braman, A. Gupta, V. Velcheti, A. Madabhushi, Predicting cancer outcomes with radiomics and artificial intelligence in radiology, Nat. Rev. Clin. Oncol. 19 (2) (2022) 132–146.

[2] J. Ulenaers, The impact of artificial intelligence on the right to a fair trial: Towards a robot judge? Asian J. Law Econom. 11 (2) (2020).

[3] Z. Zhang, H. Ning, F. Shi, F. Farha, Y. Xu, J. Xu, F. Zhang, K.-K.R. Choo, Artificial intelligence in cyber security: research advances, challenges, and opportunities, Artif. Intell. Rev. (2022) 1–25.

[4] J. Gerlings, A. Shollo, I. Constantiou, Reviewing the need for explainable artificial intelligence (xAI), 2020, arXiv preprint arXiv:2012.01007.

[5] V. Vivek Khanna, K. Chadaga, N. Sampathila, S. Prabhu, R. Chadaga P, D. Bhat, S. KS, Explainable artificial intelligence-driven gestational diabetes mellitus prediction using clinical and laboratory markers, Cogent Eng. 11 (1) (2024) 2330266.

[6] T.K. Bhat, K. Chadaga, N. Sampathila, S. KS, R. Chadaga, S. Umakanth, S. Prabhu, Predicting acute myocardial infarction from haematological markers utilizing machine learning and explainable artificial intelligence, Syst. Sci. Control Eng. 12 (1) (2024) 2331074.

[7] K. Chadaga, S. Prabhu, N. Sampathila, R. Chadaga, S. Umakanth, D. Bhat, S.K. GS, Explainable artificial intelligence approaches for COVID-19 prognosis prediction using clinical markers, Sci. Rep. 14 (1) (2024) 1783.

[8] D. Gunning, Explainable Artificial Intelligence (Xai), Defense Advanced Research Projects Agency (DARPA), 2017, nd Web.

[9] D. Das, S. Chernova, Leveraging rationales to improve human task performance, in: Proceedings of the 25th International Conference on Intelligent User Interfaces, 2020, pp. 510–518.

[10] A. Jacovi, A. Marasović, T. Miller, Y. Goldberg, Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 624–635.

[11] M. Chromik, A. Butz, Human-XAI interaction: a review and design principles for explanation user interfaces, in: Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part II 18, Springer, 2021, pp. 619–640.

[12] K. Stöger, D. Schneeberger, A. Holzinger, Medical artificial intelligence: the European legal perspective, Commun. ACM 64 (11) (2021) 34–36.

[13] U. Ehsan, Q.V. Liao, M. Muller, M.O. Riedl, J.D. Weisz, Expanding explainability: Towards social transparency in ai systems, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–19.

[14] C. Panigutti, R. Hamon, I. Hupont, D. Fernandez Llorca, D. Fano Yela, H. Junklewitz, S. Scalzo, G. Mazzini, I. Sanchez, J. Soler Garrido, E. Gomez, The role of explainable AI in the context of the AI act, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1139–1150, http://dx.doi.org/10.1145/3593013.3594069.

[15] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.

[16] M. Caro-Martínez, G. Jiménez-Díaz, J.A. Recio-García, Conceptual modeling of explainable recommender systems: an ontological formalization to guide their design and development, J. Artificial Intelligence Res. 71 (2021) 557–589.

[17] J.M. Darias, M. Caro-Martínez, B. Díaz-Agudo, J.A. Recio-Garcia, Using case-based reasoning for capturing expert knowledge on explanation methods, in: International Conference on Case-Based Reasoning, Springer, 2022, pp. 3–17.

[18] K. Martin, A. Wijekoon, N. Wiratunga, C. Palihawadana, I. Nkisi-Orji, D. Corsar, B. Díaz-Agudo, J.A. Recio-García, M. Caro-Martínez, D. Bridge, et al., iSee: intelligent sharing of explanation experiences, in: CEUR Workshop Proceedings, 2022.

[19] A. Wijekoon, N. Wiratunga, C. Palihawadana, I. Nkisi-Orji, D. Corsar, K. Martin, iSee: Intelligent sharing of explanation experience by users for users, in: Companion Proceedings of the 28th International Conference on Intelligent User Interfaces, 2023, pp. 79–82.

[20] A. Aamodt, E. Plaza, Case-based reasoning: Foundational issues, methodological variations, and system approaches, AI Commun. 7 (1) (1994) 39–59.

[21] J. Klaise, A.V. Looveren, G. Vacanti, A. Coca, Alibi explain: Algorithms for explaining machine learning models, J. Mach. Learn. Res. 22 (181) (2021) 1–7, URL http://jmlr.org/papers/v22/21-0017.html.

[22] H. Baniecki, W. Kretowicz, P. Piatyszek, J. Wisniewski, P. Biecek, dalex: Responsible machine learning with interactive explainability and fairness in python, J. Mach. Learn. Res. 22 (214) (2021) 1–7, URL http://jmlr.org/papers/v22/20-1473.html.

[23] T. Fel, L. Hervier, D. Vigouroux, A. Poche, J. Plakoo, R. Cadene, M. Chalvidal, J. Colin, T. Boissin, L. Bethune, A. Picard, C. Nicodeme, L. Gardes, G. Flandin, T. Serre, Xplique: A deep learning explainability toolbox, in: Workshop on Explainable Artificial Intelligence for Computer Vision, CVPR, 2022.

[24] J.M. Darias, B. Díaz-Agudo, J.A. Recio-Garcia, A systematic review on model-agnostic XAI libraries, in: ICCBR Workshops, 2021, pp. 28–39.

[25] C.O. Retzlaff, A. Angerschmid, A. Saranti, D. Schneeberger, R. Roettger, H. Mueller, A. Holzinger, Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists, Cogn. Syst. Res. 86 (2024) 101243.

[26] O. Dijk, oegesam, R. Bell, Lily, Simon-Free, B. Serna, rajgupt, yanhong-zhao ef, A. Gädke, A. Todor, Evgeniy, Hugo, M. Haizad, T. Okumus, woochan jang, oegedijk/explainerdashboard: explainerdashboard 0.4.2: dtreeviz v2 compatiblity, 2023, http://dx.doi.org/10.5281/zenodo.7633294.

[27] A. Wijekoon, N. Wiratunga, A user-centred evaluation of DisCERN: Discovering counterfactuals for code vulnerability detection and correction, Knowl.-Based Syst. 278 (2023) 110830, http://dx.doi.org/10.1016/J.KNOSYS.2023.110830.

[28] B. Bayrak, K. Bach, PertCF: A perturbation-based counterfactual generation approach, in: Artificial Intelligence XL: 43rd SGAI International Conference on Artificial Intelligence, AI 2023, Cambridge, UK, December 12–14, 2023, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2023, pp. 174–187, http://dx.doi.org/10.1007/978-3-031-47994-6_13.

[29] C.E. Sosa-Espadas, M. Cetina-Aguilar, J.A. Soladrero, J.M. Darias, E.E. Brito-Borges, N.L. Cuevas-Cuevas, M.G. Orozco-del-Castillo, IREX: a reusable process for the iterative refinement and explanation of classification models, in: P. Reuss, J.M. Schönborn (Eds.), Workshop Proceedings of the 30th International Conferece on Case-Based Reasoning Co-Located with the 30th International Conference on Case-Based Reasoning (ICCBR 2022), Nancy (France), September 12-15th, 2022, in: CEUR Workshop Proceedings, vol. 3389, CEUR-WS.org, 2022,

pp. 241–245, URL https://ceur-ws.org/Vol-3389/ICCBR_2022_XCBR_Challenge_ITMeridaCuevas.pdf.

[30] C. Pirie, N. Wiratunga, A. Wijekoon, C.F. Moreno-García, AGREE: a feature attribution aggregation framework to address explainer disagreements with alignment metrics, in: L. Malburg, D. Verma (Eds.), Proceedings of the Workshops At the 31st International Conference on Case-Based Reasoning (ICCBR-WS 2023) Co-Located with the 31st International Conference on Case-Based Reasoning (ICCBR 2023), Aberdeen, Scotland, UK, July 17, 2023, in: CEUR Workshop Proceedings, vol. 3438, CEUR-WS.org, 2023, pp. 184–199, URL https://ceur-ws.org/Vol-3438/paper_14.pdf.

[31] M.F. Valdez-Ávila, C. Bermejo-Sabbagh, B. Díaz-Agudo, M.G. Orozco-del-Castillo, J.A. Recio-García, CBR-fox: A case-based explanation method for time series forecasting models, in: S. Massie, S. Chakraborti (Eds.), Case-Based Reasoning Research and Development - 31st International Conference, ICCBR 2023, Aberdeen, UK, July 17-20, 2023, Proceedings, in: Lecture Notes in Computer Science, vol. 14141, Springer, 2023, pp. 200–214, http://dx.doi.org/10.1007/978-3-031-40177-0_13.

[32] R.R. Hoffman, S.T. Mueller, G. Klein, J. Litman, Measures for explainable AI: explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance, Front. Comput. Sci. 5 (2023) http://dx.doi.org/10.3389/FCOMP.2023.1096257.

[33] S. Mohseni, N. Zarei, E.D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable AI systems, ACM Trans. Interact. Intell. Syst. (TiiS) 11 (3–4) (2021) 1–45.

[34] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai, ACM Comput. Surv. 55 (13s) (2023) 1–42.

[35] B.J. Strasser, J. Baudry, D. Mahr, G. Sanchez, E. Tancoigne, "Citizen science"? Rethinking science and public participation, Sci. Technol. Stud. 32 (2) (2019) 52–76, http://dx.doi.org/10.23987/sts.60425, URL https://sciencetechnologystudies.journal.fi/article/view/60425.

[36] K. Clarkson, A.L. Gentile, D. Gruhl, P. Ristoski, J. Terdiman, S. Welch, User-centric ontology population, in: A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, M. Alam (Eds.), The Semantic Web, Springer International Publishing, Cham, 2018, pp. 112–127.

[37] I.D. Watson, Case-based reasoning is a methodology not a technology, Knowl.-Based Syst. 12 (5–6) (1999) 303–308, http://dx.doi.org/10.1016/S0950-7051(99)00020-9.

[38] M. Iovino, E. Scukins, J. Styrud, P. Ögren, C. Smith, A survey of behavior trees in robotics and AI, Robot. Auton. Syst. 154 (2022) 104096, http://dx.doi.org/10.1016/j.robot.2022.104096, URL https://www.sciencedirect.com/science/article/pii/S0921889022000513.

[39] I. Nkisi-Orji, N. Wiratunga, C. Palihawadana, J.A. Recio-García, D. Corsar, Clood cbr: Towards microservices oriented case-based reasoning, in: International Conference on Case-Based Reasoning, Springer, 2020, pp. 129–143.

[40] A. Holzinger, M. Kargl, B. Kipperer, P. Regitnig, M. Plass, H. Müller, Personas for artificial intelligence (AI) an open source toolbox, IEEE Access 10 (2022) 23732–23747.

[41] G. Flórez-Puga, P.A. González-Calero, G. Jiménez-Díaz, B. Díaz-Agudo, Supporting sketch-based retrieval from a library of reusable behaviours, Expert Syst. Appl. 40 (2) (2013) 531–542, http://dx.doi.org/10.1016/j.eswa.2012.07.067, URL https://www.sciencedirect.com/science/article/pii/S0957417412009232.

[42] A. Wijekoon, D. Corsar, N. Wiratunga, Behaviour trees for creating conversational explanation experiences, 2023, arXiv:2211.06402.

[43] I. Nkisi-Orji, C. Palihawadana, N. Wiratunga, D. Corsar, A. Wijekoon, Adapting semantic similarity methods for case-based reasoning in the cloud, in: International Conference on Case-Based Reasoning, Springer, 2022, pp. 125–139.

[44] A. Wijekoon, N. Wiratunga, K. Martin, D. Corsar, I. Nkisi-Orji, C. Palihawadana, D. Bridge, P. Pradeep, B.D. Agudo, M. Caro-Martínez, CBR driven interactive explainable AI, in: International Conference on Case-Based Reasoning, Springer, 2023, pp. 169–184.

[45] M. Caro-Martinez, J.M. Darias, B. Diaz-Agudo, J.A. Recio-Garcia, iSeeE3—The explanation experiences editor, SoftwareX 21 (2023) 101311.

[46] B. Paaß en, B. Mokbel, B. Hammer, A toolbox for adaptive sequence dissimilarity measures for intelligent tutoring systems, in: O.C. Santos, J.G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J.M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, M. Desmarais (Eds.), Proceedings of the 8th International Conference on Educational Data Mining, EDM 2015, International Educational Datamining Society, 2015-06, 632–632.

[47] V.I. Levenshtein, et al., Binary codes capable of correcting deletions, insertions, and reversals, in: Soviet Physics Doklady, Vol. 10, Soviet Union, 1966, pp. 707–710.

[48] A. Wijekoon, N. Wiratunga, D. Corsar, K. Martin, I. Nkisi-Orji, B. Díaz-Agudo, D. Bridge, XEQ scale for evaluating XAI experience quality grounded in psychometric theory, 2024, arXiv:2407.10662. URL https://arxiv.org/abs/2407.10662.

[49] R. Shokri, M. Strobel, Y. Zick, On the privacy risks of model explanations, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021, pp. 231–241.

[50] K. Balog, F. Radlinski, Measuring recommendation explanation quality: The conflicting goals of explanations, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 329–338.

[51] H. Chen, Y. Zhang, Y. Cao, J. Xie, Security issues and defensive approaches in deep learning frameworks, Tsinghua Sci. Technol. 26 (6) (2021) 894–905.

[52] E. Fink, M. Sharifi, J.G. Carbonell, Application of machine learning and crowdsourcing to detection of cybersecurity threats, in: Proceedings of the US Department of Homeland Security Science Conference–Fifth Annual University Network Summit, Washington, DC, 2011.

[53] C. Cath, S. Wachter, B. Mittelstadt, M. Taddeo, L. Floridi, Artificial intelligence and the 'good society': the US, EU, and UK approach, Sci. Eng. Ethics 24 (2018) 505–528.