WIJEKOON, A., WIRATUNGA, N., CORSAR, D., MARTIN, K., NKISI-ORJI, I., PALIHAWADANA, C., CARO-MARTÍNEZ, M., DÍAZ-AGUDO, B., BRIDGE, D. and LIRET, A. 2024. iSee: advancing multi-shot explainable AI using case-based recommendations. In *Endriss, U., Melo, F.S., Bach, K., et al. (eds.) ECAI 2024: proceedings of the 27th European conference on artificial intelligence, co-located with the 13th conference on Prestigious applications of intelligent systems (PAIS 2024), 19–24 October 2024, Santiago de Compostela, Spain*. Frontiers in artificial intelligence and applications, 392. Amsterdam: IOS Press [online], pages 4626-4633. Available from: https://doi.org/10.3233/FAIA241057

iSee: advancing multi-shot explainable AI using case-based recommendations.

WIJEKOON, A., WIRATUNGA, N., CORSAR, D., MARTIN, K., NKISI-ORJI, I., PALIHAWADANA, C., CARO-MARTÍNEZ, M., DÍAZ-AGUDO, B., BRIDGE, D. and LIRET, A.

2024

ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA241057

iSee: Advancing Multi-Shot Explainable AI Using Case-Based Recommendations

Anjana Wijekoon[a], Nirmalie Wiratunga[a,*], David Corsar[a], Kyle Martin[a], Ikechukwu Nkisi-Orji[a], Chamath Palihawadana[a], Marta Caro-Martínez[b], Belen Díaz-Agudo[b], Derek Bridge[c] and Anne Liret[d]

[a]Robert Gordon University, Aberdeen, Scotland
[b]Universidad Complutense de Madrid, Spain
[c]University College Cork, Ireland
[d]British Telecommunications, France

Abstract. Explainable AI (XAI) can greatly enhance user trust and satisfaction in AI-assisted decision-making processes. Recent findings suggest that a single explainer may not meet the diverse needs of multiple users in an AI system; indeed, even individual users may require multiple explanations. This highlights the necessity for a "multi-shot" approach, employing a combination of explainers to form what we introduce as an "explanation strategy". Tailored to a specific user or a user group, an "explanation experience" describes interactions with personalised strategies designed to enhance their AI decision-making processes. The iSee platform is designed for the intelligent sharing and reuse of explanation experiences, using Case-based Reasoning to advance best practices in XAI. The platform provides tools that enable AI system designers, i.e. design users, to design and iteratively revise the most suitable explanation strategy for their AI system to satisfy end-user needs. All knowledge generated within the iSee platform is formalised by the iSee ontology for interoperability. We use a summative mixed methods study protocol to evaluate the usability and utility of the iSee platform with six design users across varying levels of AI and XAI expertise. Our findings confirm that the iSee platform effectively generalises across applications and its potential to promote the adoption of XAI best practices.

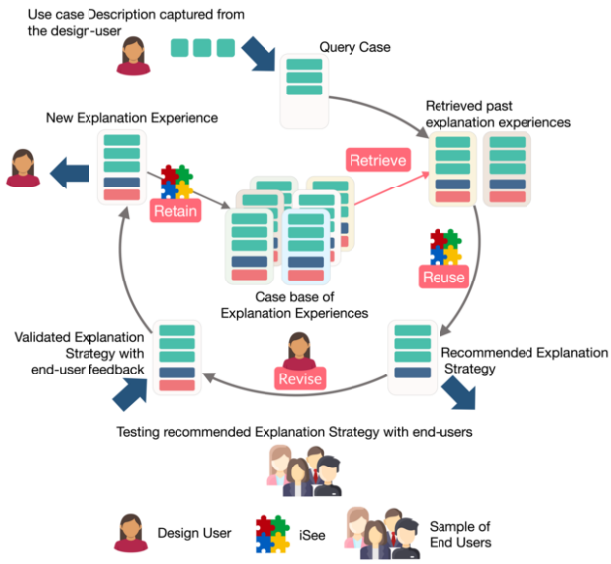

Figure 1. iSee CBR Cycle

1 Introduction

XAI systems must be able to address a range of explanation needs (such as transparency, scrutability, and trust) and must do so in a manner that is relevant to a range of stakeholders. It is also now a regulatory requirement in many parts of the world such as the right to obtain an explanation in the EU [8, 13]. It is essential for an AI system looking to implement XAI to learn from successful past experiences of XAI adaptations that reveal best practices. Case-based Reasoning (CBR) caters to this need whereby it learns from past experiences [2, 14]. The iSee platform has proposed utilising the CBR paradigm to capture knowledge and experience from successful adaptations of explainability within AI systems [37].

It is increasingly recognised that a single explanation is often insufficient to satisfy all situations and/or all stakeholders [3, 25]. Multi-shot explanations, allowing users to digest explanations from multiple algorithms over the course of a single interaction, have been demonstrated to provide more satisfactory user experiences [24, 37]. However, bespoke development of domain-specific and personalised explanation strategies is prohibitively expensive and requires significant domain and XAI expertise. Explainability toolkits have emerged to facilitate the development of explainability for use cases, including IBM-XAI 360 [4], Alibi [19] and Captum [20]. Where existing toolkits fall short, iSee fills this gap by providing support to design users, regardless of their level of expertise in XAI, to develop explanation strategies based on the collective experiences of others.

The iSee Cockpit is designed to elicit stakeholder requirements for explainability from the design user, which drives the underpinning CBR paradigm of the platform. Accordingly, in this paper, we evaluate the iSee Cockpit tools for capturing these requirements from real-world design users towards the design of multi-shot explanation strategies. We make two key contributions: 1) A formalisation of the multi-shot explanation experience underpinned by the CBR paradigm; and 2) The design and findings of a user experience evaluation of the iSee cockpit with insights from six design users, highlighting the tool's usability, utility, and areas for improvement.

The rest of this paper is organised as follows. Related literature is

* Corresponding Author. Email: n.wiratunga@rgu.ac.uk

discussed in Section 2. Section 3 presents the iSee platform, and contextualises the system using a radiography fracture detection AI system. Section 4 presents the user experience study design and results are presented in Section 5, followed by conclusions in Section 6.

2 Related Work

CBR as a methodology is naturally interpretable due to its central role in harnessing episodic knowledge in the form of cases, where the assumption that cases of similar problems have similar solutions is innately easy for end users to understand and act upon. The link between XAI and CBR is not new, with some of the earliest works dating back to the eighties, where the focus was on explanation methods to support further interpretability in CBR i.e. XAI for CBR. Here good explanations were considered as those that took into account the explainer's goals and beliefs.

The need for explanations is specific to each user, and assisting them in expressing this need requires personalisation, considering both situational and individual elements, as highlighted by Leake [21]. Typically explanations were represented as part of the case structure using knowledge-rich rules and scripts to store explanation knowledge [30]. Explanation-based indexing commonly used features known to be good at explaining the cause of faults as indices for case retrieval [17, 6, 1]. In this way, indexing knowledge could also be used effectively to explain case retrieval and improve case ranking by combining causal knowledge. In order to reduce the burden of gathering explanations at case creation, research began to focus on reusing past case-based explanations and manually adapting them to fit current anomalous situations needing explanation [22, 31]. This is sensible as similar problem situations are likely to benefit from similar explanations. This idea is similar to our iSee platform, which also exploits past explanation experiences for new situations. However, unlike these past works, explainers in iSee are aimed at explaining AI black box models, not CBR systems.

In recent work methods from CBR, specifically CBR's similarity knowledge, have been employed to generate factual, counterfactual, and semi-factual explanations for black-box models. Factual k-NN Explanations use model-specific neural prototypes [23] and model-agnostic twin systems [18] to provide clear and understandable justifications for AI decisions. In some scenarios, neighbours within the case base act as explainers [12], providing locally faithful surrogates or twin explanations [28]. Counterfactual k-NN Explanations also use similarity knowledge but to identify Nearest Unlike Neighbours (NUNs) for valid action recommendations helping users understand what minimal changes could alter an AI's decision [40, 33, 7]. Semi-Factual k-NN Explanations combine factual and non-factual explanations (with Farthest Like Neighbours and NUN combinations) to offer a more comprehensive understanding of the decision-making process [35, 5].

The adaptation step in CBR commonly includes constructive [29] and transformative [10] adaptation. Both of which we make use of within iSee to enable the reuse of past explanation experiences having tailored them to the current situation. Our work is unique in that we introduce adaptation operators applicable for explainer method reuse. The need for interactive XAI methodologies is closely linked to aligning with the evolving mental models of end users [15]. iSee also employs a dialogue interface to facilitate interaction, guided by transitions prescribed by the explanation strategy. This ensures that explanation strategies are both effective and user-friendly.

3 iSee Platform

The goal of the iSee platform is to help design users create and refine explanation strategies for their XAI systems to ensure end-user satisfaction. Using the CBR 4R steps, iSee is organised to retrieve, reuse, revise, and retain explanation experiences as cases. Figure 1 illustrates how the iSee platform is underpinned by the CBR paradigm. Central to CBR's 4Rs are its knowledge containers: case base, case similarity and case adaptation. In iSee, these containers are formalised using the iSee ontology for interoperability.

3.1 iSee platform overview

The iSee Cockpit elicits explainability requirements from a design-user, who is an expert of the AI system's design and its stakeholder needs. These requirements form the query to our case base of past experiences, facilitating the retrieval of the most suitable explanation strategies. iSee provides tools to automatically adapt a recommended solution to further match design-user requirements by reusing multiple explanation strategies from nearest neighbours. The design user can then evaluate a recommended (and adapted) strategy solution with a representative sample of their stakeholders to get feedback that can then be used for collaborative revision of the case description and solution. Once the stakeholder explanation needs are met, the design user can finalise the validated explanation strategy for their AI system thus forming a new case. The quality and coverage of cases in the case base enhances case-based recommendations. Accordingly retaining a complete anonymised case with the design user's consent is an important last step in iSee's 4R CBR cycle.

The iSee platform was implemented using a micro-service based approach. Each module of the platform (i.e. user interface, case retrieval, failure-driven adaptation, etc) can be hosted and executed independently on a single or multiple servers. Modules are logically connected to each other through standardised API endpoints, allowing flexibility for allocation of computational resources required to execute them.

3.2 Explanation Experience Case Base

An explanation experience case is a multi-faceted entity that encapsulates several knowledge constructs: the attributes of the AI system; user groups and their explanation needs; the explanation strategy; and user explanation experience feedback (see Figure 2). More formally the iSee case base is a collection of past explanation experiences, each case c represented as a triplet.

$$c = \{c_{\mathcal{D}}, c_{\mathcal{S}}, c_{\mathcal{O}}\}$$

Where case description ($\mathcal{D}$) covers the constructs related to explanation requirements, a solution ($\mathcal{S}$) representing the explanation strategy and an outcome $\mathcal{O}$ capturing user feedback. Here a query q is a case where the solution and the outcome are empty ($\mathcal{S}, \mathcal{O} = \emptyset$). The majority of the cases are selected from literature following a critical review and we include several anonymised industry cases. iSee utilises these cases to recommend strategies to design users who are looking to build explainability in their AI systems.

An explanation strategy, i.e, the case solution is modelled using a Behaviour Tree (BT) [9]. The example explanation strategy BT in Figure 3 is executed as follows. If the user asks a "why" kind of question, answer them with a GradCAM explanation and if they need to verify with another type of explanation (variant) provide the nearest neighbours; if the user is still not satisfied and asks a further "what"

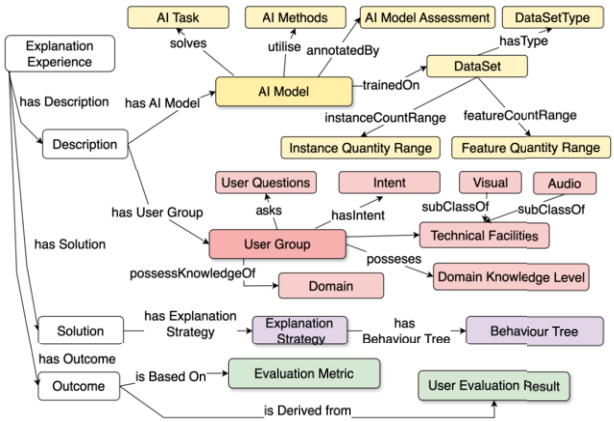

Figure 2. Explanation Experience Case

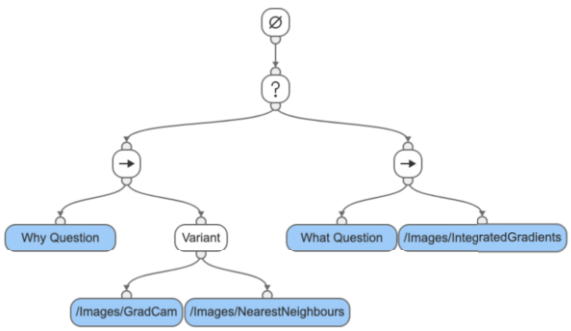

Figure 3. Sample Case Solution

kind of question, provide them an Integrated Gradients explanation. In this way, a BT model is a good way to manage execution in a controlled manner as it allows systematic handling of different types of user questions and corresponding explanation strategy sub-trees.

3.3 Case Retrieval using Similarity Knowledge

The retrieval task finds similar cases from the case base using similarity knowledge, which specifies local similarity metrics and aggregates them into a global similarity score.

$$local\_sim = \begin{cases} WP & \text{if } j \in [\textit{AI Task}, \textit{AI Method}] \\ QI & \text{if } j \in [\textit{Technical Facilities}, \textit{User Questions}] \\ EM & \text{otherwise} \end{cases}$$

where j refers to a feature in $\mathcal{D}$. The similarity metrics are as follows.

Wu & Palmer (WP) is a taxonomy path-based similarity metric originally implemented for calculating word similarities. For *AI Task* and *AI Method* case attributes, given the taxonomic representation from iSee ontology, it computes node similarity between the query and the case nodes based on node depths and distances from the most specific common ancestor [26].

Query Intersection (QI) is applicable for attributes where the data type is a set of ontology individuals such as attributes *User Questions* and *Technical Facilities*. Given the feature j from query q and case i, it calculates the similarity as the intersection between two sets normalised by the query set size as $(|c_j^q \cap c_j^i|)/|c_j^q|$.

Exact Match (EM) similarity indicates a string match. This is applied both for case attributes that are ontology individuals, and is the most common method of comparison.

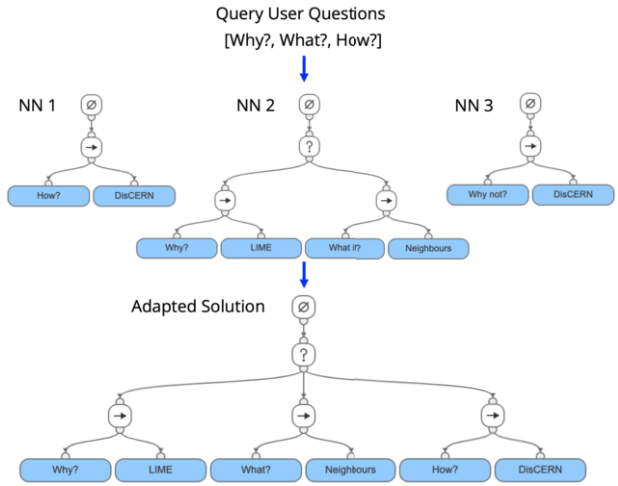

Figure 4. Failure Driven Adaptation

iSee implements retrieval using CloodCBR [26] in two phases: 1) filter case base to only include cases that exactly match the query *DataSetType* (dt) (Equation 1); and 2) calculate pair-wise similarity to each filtered case to select the top k most similar cases (Equation 2).

$$\mathcal{C}' = \{c^i \in \mathcal{C} \mid c^q_{dt} = c^i_{dt}\} \tag{1}$$

$$\text{Top-}k = \underset{c^i \in \mathcal{C}'}{\text{argmax}^k} \left\{ \frac{1}{|\mathcal{D}|} \sum_{j=1}^{|\mathcal{D}|} local\_sim(c^q_j, c^i_j) \right\} \tag{2}$$

The similarity between the query case c^q, and a case c^i from the case subset $\mathcal{C}'$ is calculated as the aggregation of local similarities as in Equation 2. The single top case, is the case with the highest global similarity score among the top k and the recommended explanation strategy for the query requirements.

3.4 Explanation Strategy Reuse

The CBR methodology recommends solution adaptation before reuse to: 1) address unmet requirements on the query description; and 2) personalise the solution utilising domain knowledge. iSee offers a failure-driven adaptation algorithm to address the former and envision the latter to be a manual process.

Adaptation of solutions is driven by the failure to fulfil the query's *User Questions*. The mismatch between the recommended case and the query is calculated using the Query Intersection similarity and when similarity is ≤ 1 we apply a stable marriage algorithm on the top k case solutions ($k > 1$) to find user question-explanation strategy sub-trees that satisfy all (or as many) of the user questions that appear in the user's query. The *adapted* explanation strategy BT is formed of these sub-trees. Figure 4 presents an example where 2 of the query user questions ("Why" and "What") are not met by the recommended case solution (NN1). Accordingly, iSee uses the top 3 neighbours to find best sub-tree matches in neighbours 2 and 3 to form the adapted solution [27].

3.5 Explanation Strategy Revision

iSee provides an editor and the following supporting tools for design users to revise explanation strategies.

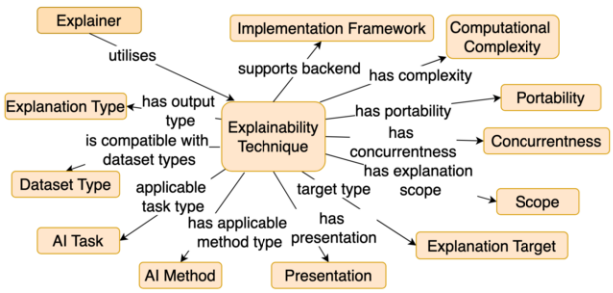

Figure 5. Explainer Properties

Explainer Applicability warns the design user about the implementation mismatches between the explainers in the recommended strategy and their query case. These mismatches include 1) the implementation framework supported by the explainers and that of the query AI model; 2) model access requirements of the explainers and model access provided by the design user (model file or API access to the predict function); and 3) labelled data requirements of the explainers and data provided by the design user.

Explainer Substitution provides the design user with substitution recommendations for a selected explainer. This follows a similar approach to retrieval on a library of explainers where each is characterised using a set of semantic features (see Figure 5). The similarity between explainers is calculated as $e\_sim(e^q, e^i) = \frac{1}{|\mathcal{M}|}\sum_{j=1}^{|\mathcal{M}|} local\_e\_sim(e^q_j, e^i_j)$ using the following local similarities.

$$local\_e\_sim = \begin{cases} WP & \text{if } j \in [\textit{AI Tasks}, \textit{AI Methods}, \textit{Presentation}] \\ QI & \text{if } j \in [\textit{Implementation Frameworks}, \\ & \textit{Explanation Technique}, \textit{Explanation Type}] \\ EM & \text{otherwise} \end{cases}$$

Sub-tree Substitution provides applicable sub-tree substitutions for a selected sub-tree. Substitutions are selected from the solutions of similar cases based on edit-distance similarity. iSee transforms the query and case sub-trees into directed graphs and calculates edit distance using the node similarities defined below where $type(n)$ returns the strategy node type.

$$node\_sim = \begin{cases} e\_sim & \text{if } type(\{n^q, n^i\}) = \text{Explainer} \\ sem\_sim(.) & \text{if } type(\{n^q, n^i\}) = \text{User Question} \\ 1 & type(n^q) = type(n^i) \\ 0 & type(n^q) \neq type(n^i) \end{cases}$$

3.6 Case Retention

Once the design user has a recommended solution, adapted and/or revised, they evaluate it with their stakeholders. iSee provides a chatbot interface where stakeholders can execute the strategy to create explanation experiences and provide feedback [36, 38]. Design users can utilise this feedback to iteratively improve the case description and the solution, aiming for stakeholder satisfaction.

We form the case outcome from the stakeholder feedback obtained using the XAI Experience Quality (XEQ) Scale [39] tool which was developed as a psychometric scale for measuring the quality of explanation experiences across 4 dimensions: Learning, Utility, Fulfilment and Engagement. The case outcome records the mean score in each dimension. During case retention, iSee creates an anonymised copy of the complete case and retains it in the case base. A case maintenance policy [11] can then be used to periodically review the case base considering case coverage.

3.7 Example Use Case: Radiograph Classification

We describe a radiograph classification system provided by an industry partner. The AI system is implemented using ConvNet-based architecture for binary classification of fractures in radiographs. The stakeholder explanation needs of this use case stem from two factors: 1) to improve the quality of their product for end-users; and 2) to increase regulatory compliance with relevant governance bodies. The design user described two user groups: 1) clinicians who are using the AI system for decision support; and 2) managers who are looking to evaluate the compliance, risk and regulatory requirements.

Using callouts of iSee screenshots in Figure 6, we illustrate how a design user can interact with the iSee *retrieve*, *reuse*, and *revise* tools to create a complete Explanation Experience case, containing both the case description and solution parts, and *retain* it in the case base. Firstly, an AI model description and implementation of a ConvNet model for binary classification of black and white radiography images is entered into the iSee Cockpit. Further details on how to access the model can also be provided. User groups and intents part of the description include details of a clinician persona, alongside corresponding intents in transparency and performance, thus completing the query case description $q_{\mathcal{D}}$. The completed case description parts can be used to query the iSee case base and *retrieve* a set of candidate cases containing previous best practices of explanation strategies. In the example in Figure 6, of the retrieved three cases containing variations of strategies include a combination of feature attribution and nearest neighbour-based explainers (top of blue callout). The design user can decide to *reuse* the recommended solution arrived at after iSee performs a failure-driven transformational adaptation to obtain a personalised strategy. After that, they can decide to perform a manual *revise* step using a strategy editor (bottom of blue callout), which will provide a ranked list of substitute explainers for any selected explainer node that the designer user wishes to change (as demonstrated here by highlighting the Integrated Gradients explainer node for substitution). Once revisions are complete, the case contains the refined solution component. It can be evaluated with target stakeholders to identify the case outcome (which is measured against the dimensions of the XEQ Scale). This allows the formation of a complete case, which can subsequently be *retained* in the case base to inform future practice.

4 User Experience Evaluation with Design Users

A summative assessment, using a mixed methods study, was performed to evaluate the user experience of design users. We aim to evaluate the following two dimensions.

- Utility: Do the design users perceive the tool as fit for purpose?
- Usability: Do the design users find it easy and efficient to complete their tasks?

4.1 Study Protocol

We planned a two-stage user-centred evaluation session with a design user lasting approximately 1 hour. A session is standardised using the following protocol:

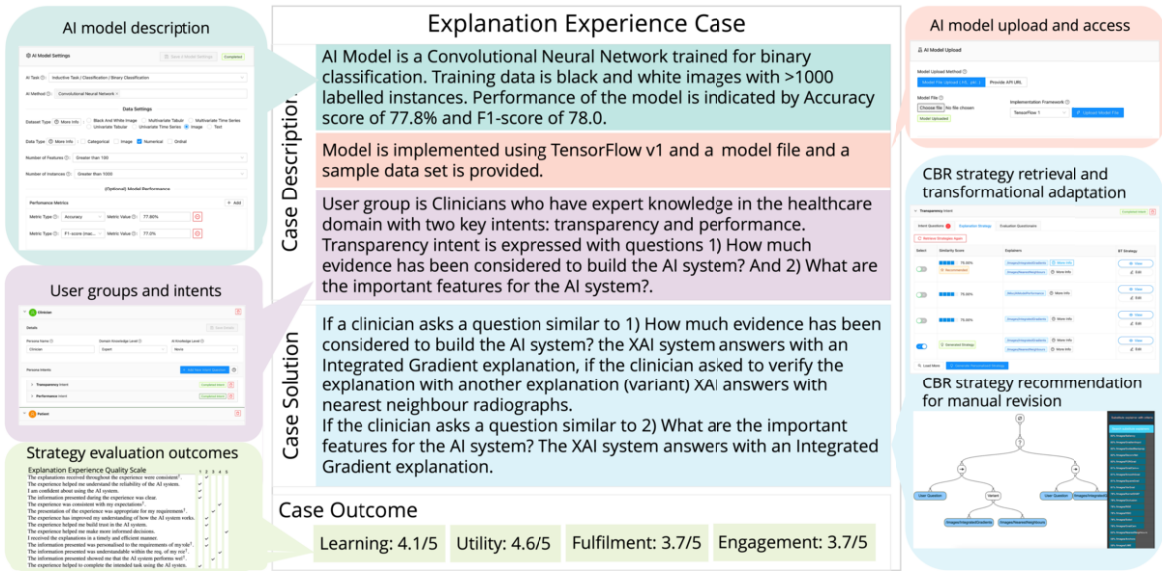

Figure 6. Radiograph Classification Use Case created following the iSee CBR 4Rs - Please refer to the supplementary material on arXiv to access the full-resolution screenshots of the platform

Table 1. Design User Profiles

ID	Participant Description	Experience in AI	Experience in XAI
D1	Academic Researcher, designed and developed the AI system	Proficient	Novice
D2	Lead for AI system delivery within the company, involved in the requirements management and design of the AI system	Expert	Novice
D3	Lead data scientist of the company involved in the requirements management, design and development of the AI system	Expert	Expert
D4	Academic Researcher, designed and led the development of the AI system	Expert	Proficient
D5	Academic Researcher, designed and led the development of the AI system	Expert	Expert
D6	Lead data scientist of the company involved in the requirements management, design and development of the AI system	Expert	Expert

1. To open the session, the researcher provided a brief overview of the iSee project and the objectives of the session. A toy example of a loan approval XAI system was used to illustrate the specific information requested on the Cockpit.
2. We then conducted a concurrent Think-Aloud Protocol (TAP) where the participant was given access to the iSee Cockpit and asked to create their use case as a design-user of the system. Throughout the session, participants were encouraged to vocalise their thoughts. The researcher intervened only when necessary, (i.e. when the user sought clarification or was unable to proceed).
3. On completion of the TAP, the participant was asked to respond to the User Experience Questionnaire (UEQ) consisting of 26 questions on a 7-step Likert scale.
4. To conclude the session, the researcher asked participants a series of open-ended questions. These questions aimed to establish a design user profile and capture any additional comments or insights regarding their experience.

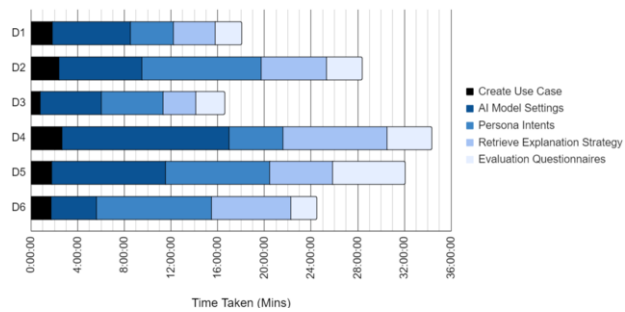

Figure 7. iSee Cockpit time taken per component

4.2 Recruitment

This study involved six design users, two conducted in person and four online over Microsoft Teams. They were recruited through existing academic and industry connections with the leading institution. These design users were distinct from any design users who were involved in the initial UI/UX design activities to avoid bias. At the start of the study, the leading researcher obtained informed consent from the design user to record the screen and audio during the session and to use the data generated during the session exclusively for research purposes. The design user profiles are summarised in Table 1. In the rest of the paper, the ID column is used to reference each design user.

4.3 Study Instruments

The UEQ questions assessed user perception of usability (UEQ dimensions efficiency, perspicuity and dependability) as well as user experience (UEQ dimensions novelty and stimulation). The overall impression of the product is measured by the attractiveness dimension. For our study, we were interested in 2 out of the 4 objectives

of the original UEQ framework related to establishing sufficient user experience and identifying areas of improvement. For a more detailed discussion on the UEQ, please refer to [32].

The UEQ benchmark is based on an analysis of 163 products consisting of business applications, development tools, e-commerce websites, social networks and mobile applications. Each dimension measures the scores in 5 categories from *excellent*, *good*, *above average*, *below average* and *bad*, and a new product is expected to reach *good* category in all dimensions. We note that the benchmark has not considered technological or research-oriented products like the iSee platform. Also, due to the limited number of use cases, we are not able to establish the statistical significance of the results. The Concurrent TAP method was used to obtain qualitative insights into user experience by asking users to verbalise their thoughts as they use a system to perform a specific task [16, 34]. These sessions facilitate targeted usability evaluation of specialist tools, as they allow flexible task performance and researcher intervention when required.

4.4 Analysis

Using the above instruments, the study generated three artefacts: 1) transcriptions of the audio and screen recordings; 2) researcher notes documented during TAP sessions; and 3) UEQ responses. They are utilised in a two-part quantitative and qualitative analysis: 1) measure user experience against established UEQ benchmarks; and 2) combine UEQ responses, the researcher notes, and transcripts to perform a thematic analysis of TAP session outcomes.

We used the recording to analyse the time taken to use the cockpit to produce a functional explanation strategy. The starting point was when participants clicked the *'Create Use Case'* button, and the endpoint was when participants saved the evaluation questionnaire (which is the final stage of use case creation). The mean time taken was 25 minutes and 51 seconds and a breakdown of time taken per component of the Cockpit is available in Figure 7.

Despite the relative freedom of the TAP session, all users converged on a similar progression through the components of the Cockpit interface. The similarity of user pathways highlights the interface is structured in a logical manner. Examining individual sessions, all 3 industry design users spent the majority of their time (%) identifying persona intents while academic design users primarily focused on configuring AI model settings. This suggests a difference in prioritisation for explanation strategy design where industry users are focused on end-user needs, while academics are focused on model details. We highlight that differences in expertise level do not seem to reflect in the time taken to complete the exercise. This promising result highlights the platform is equally suitable for expert and non-expert design users.

5 Results and Discussion

5.1 UEQ Findings

Overall, the Cockpit has been scored above *above average*, and achieves *good* category on user experience (i.e. Stimulation and Novelty). Figure 8 presents how the iSee Cockpit scored across the six dimensions measured by UEQ. The y-axis scale ranges from +3 to -3 with the mean response of 0 representing a *neutral* sentiment. For Attractiveness, Perspicuity and Efficiency, the cockpit scores *above average*, while scoring *below average* for Dependability. It is noteworthy that the benchmark has been established on products intended for the general public whereas the Cockpit caters to a specialised

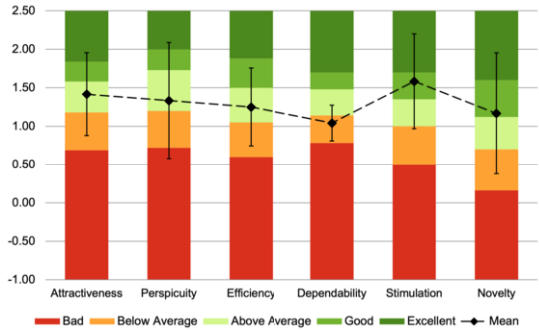

Figure 8. iSee Cockpit user experience against UEQ Benchmark

Figure 9. UEQ individual response distributions

group of expert users. Despite this, we have achieved *above average*, which is very promising.

Figure 9 presents a detailed view of individual responses. We highlight that responses are overall positive; 20/26 questions are *majority positive*, with the remaining 6 questions being *majority neutral* or equally *split* between positive and neutral. Importantly, there are no questions with majority-negative responses (with only negative or neutral responses). We explore the justification for these responses by examining the TAP session transcripts in the following section.

5.2 TAP Findings

Here we present the findings of a thematic analysis broken down across the UEQ themes and evidenced using TAP session transcripts.

Attractiveness Users had an overall positive view of the iSee Cockpit, and quickly identified the value that stakeholder-driven explanation strategies would add to their practice:

> *Each time [an AI] gives a prediction, I don't trust it at face value. So getting the the evidence behind the prediction is one of the high priorities for me.* - D3

Similar comments were made by all users. We take this as evidence the iSee Cockpit contributes to satisfying business needs for low-code development of explanation strategies. There were comments regarding the presentation of the Cockpit interface. Most comments targeted improving guidance for navigation or reducing verbiage (see below discussion of Perspicuity for examples). Otherwise, users were satisfied with the presentation.

Perspicuity All users required prompting at different stages of using the iSee system. The Cockpit contains a large volume of information, which users sometimes find complicated or intimidating (as evidenced in Figure 9 responses for complicated/easy). This was evidenced throughout the TAP sessions (*"These are not common words."* - D1; *"There is a lot here"* - D3). Despite this, we highlight that all users responded very positively to the UEQ statement the system is easy to learn. This highlights that users feel the Cockpit will be easier to use in subsequent iterations, but requires more scaffolding on first use. To resolve this, we plan to develop a set of tutorials to improve the initial experience of using the system.

Efficiency Users found that using the Cockpit was efficient. D5 observed that pre-filled components derived from the ontology facilitated data entry:

> *I really like that you have these options to kind of pick from. I think it makes that much easier.* - D5

The information from the ontology therefore provides both standardisation of data input for case formalisation, as well as user support. From observing the video recordings of TAP sessions, we identified that navigation of the interface required direction. This mostly impacted areas where the subsequent task was not clear (for example, several users did not realise that an explanation strategy was retrieved for each persona intent). We will address this with added navigation support in future versions of the system.

Dependability The iSee cockpit scored *Below Average* on the Dependability dimension, and one reason for this was security was highlighted as a key concern (highlighted by the *majority-neutral* response to not secure/secure in Figure 9). Three use cases in this study were developed as open-science and the others were AI systems developed for proprietary use. Design users from both categories raised questions about security and privacy. For example, D2 highlighted the need for authenticating access to model APIs, and D1 requested confirmation of web page secuity (*"is it HTTPS?"*) before providing any input. Despite this, design users expressed satisfaction over features such as encrypted network communication and authentication provided by the Cockpit.

Stimulation As evidenced in Figure 9, all users found the Cockpit highly motivating. D4 highlighted that the capability of iSee to facilitate comparison of multiple explainers would be very useful for testing, validating and understanding the model:

> *If I had a tool like this, it would be really helpful. When you develop [explanations], sometimes you have this kind of error analysis and if you have this kind of tool where you have several explanations, it's really, really interesting. Sometimes I had only one explainer and sometimes it doesn't work or it is biased, or focuses on things that are not really relevant.* - D4

Similarly, D1 and D2 highlighted that the interaction had motivated them to test the tool with actual users:

> *Yeah, that's fantastic. I have a really nice explanation. As an AI developer I can create my personas and such that I can then get feedback on the particular explainers.* - D2

We highlight these factors as evidence of the iSee system's utility, as inspiring users to develop and refine their explanation strategies is a key feature of the tool.

Novelty Feedback from users emphasised they found the novel aspects engaging. Specifically, users highlighted that the wealth of different explainers was exciting:

> *You have a lot of techniques there, which I'm going to have to go and have a look at.* - D2

Additionally, users expressed an interest in the underpinning methodology. D6 engaged in a discussion regarding CBR and the opportunities for empowering different explanations at an instance level (i.e. explaining complex instances using different explainer algorithms from simple instances). This evidences that the iSee platform encourages creative thinking and supports the sharing of best practices.

Overall, we found the results of the TAP sessions to be positive with encouragement for improvements. User comments actively supported the UEQ responses and highlighted user experience of the iSee Cockpit was equally satisfying for expert and non-expert users across use cases. Improving the usability of the interface shall be a key target in the ongoing development. Outcomes of the evaluation emphasise the need for clearer navigation support to facilitate interaction within the Cockpit. Finally, the perceived security of the interface will be improved, scoring better on the Dependability dimension. We will address these as part of ongoing development in the iSee system.

6 Conclusion

In this paper, we have described the implementation and evaluation of the iSee system. The iSee platform is based on CBR for the reuse of best practices in creating multi-shot explanation experiences. We presented the findings of a comprehensive user-centred evaluation including both industry and academic participants where we demonstrated the utility and usability of the system via the recognised UEQ benchmark, and analysis of think-aloud session outcomes. Our findings highlighted that both expert and non-expert design users found iSee comparably useful in assisting the implementation of multi-shot XAI in their AI systems. In future, the iSee platform will continue to grow by evaluating and improving the iSee Cockpit to enhance the design user experience; improving the coverage of the case base for improved recommendations and extending the availability of explanation methods for improved adaptation and revision.

Ethical Statement

The study protocol was reviewed and approved by the lead institution's ethics review committee. Informed consent was obtained from all participants.

Acknowledgements

The authors would like to thank Jiva.ai for providing the radiograph classification use case described in this paper, and all design users who participated in the user experience evaluation.

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; for France from The French National Research Agency under grant number ANR-21-CHR4-0004 and for Spain from the MCIN/AEI and European Union "NextGenerationEU/PRTR" under grant number PCI2020-120720-2.

References

[1] A. Aamodt. Explanation-driven case-based reasoning. In *European Workshop on Case-Based Reasoning*, pages 274–288. Springer, 1993.

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

[3] V. Arya, R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, et al. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012*, 2019.

[4] V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. MojsiloviÄ‡, S. Mourad, P. Pedemonte, R. Raghavendra, J. T. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, and Y. Zhang. Ai explainability 360: An extensible toolkit for understanding data and machine learning models. *Journal of Machine Learning Research*, 21(130):1–6, 2020. URL http://jmlr.org/papers/v21/19-1035.html.

[5] S. Aryal and M. T. Keane. Even-ifs from if-onlys: Are the best semi-factual explanations found using counterfactuals as guides? In *International Conference on Case-Based Reasoning*, pages 33–49. Springer, 2024.

[6] R. Barletta and W. Mark. Explanation-based indexing of cases. In *AAAI*, pages 541–546, 1988.

[7] D. Brughmans, P. Leyman, and D. Martens. Nice: an algorithm for nearest instance counterfactual explanations. *Data mining and knowledge discovery*, pages 1–39, 2023.

[8] C. Cath, S. Wachter, B. Mittelstadt, M. Taddeo, and L. Floridi. Artificial intelligence and the 'good society': the us, eu, and uk approach. *Science and engineering ethics*, 24:505–528, 2018.

[9] M. Colledanchise and P. Ögren. *Behavior trees in robotics and AI: An introduction*. CRC Press, 2018.

[10] S. Craw, N. Wiratunga, and R. C. Rowe. Learning adaptation knowledge to improve case-based reasoning. *Artificial intelligence*, 170(16-17): 1175–1192, 2006.

[11] S. Craw, S. Massie, and N. Wiratunga. Informed case base maintenance: A complexity profiling approach. In *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*, volume 22, page 1618. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007.

[12] D. Doyle, P. Cunningham, D. Bridge, and Y. Rahman. Explanation oriented retrieval. In *Advances in Case-Based Reasoning: 7th European Conference, ECCBR 2004, Madrid, Spain, August 30-September 2, 2004. Proceedings 7*, pages 157–168. Springer, 2004.

[13] European Parliament and Council. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), 2016.

[14] K. J. Hammond. Case-based planning: A framework for planning from experience. *Cognitive science*, 14(3):385–443, 1990.

[15] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman. Measures for explainable ai: explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance. *Front. Comput. Sci. 5, 1096257 (2023)*, 2023.

[16] R. Jääskeläinen. Think-aloud protocol. *Handbook of translation studies*, 1:371–374, 2010.

[17] A. Kass, D. Leake, and C. Owens. Swale: A program that explains. *Explanation patterns: Understanding mechanically and creatively*, 1(1): 232–254, 1986.

[18] E. M. Kenny and M. T. Keane. Twin-systems to explain artificial neural networks using case-based reasoning: Comparative tests of feature-weighting methods in ann-cbr twins for xai. In *Twenty-Eighth International Joint Conferences on Artifical Intelligence (IJCAI), Macao, 10-16 August 2019*, pages 2708–2715, 2019.

[19] J. Klaise, A. V. Looveren, G. Vacanti, and A. Coca. Alibi explain: Algorithms for explaining machine learning models. *Journal of Machine Learning Research*, 22(181):1–7, 2021. URL http://jmlr.org/papers/v22/21-0017.html.

[20] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020.

[21] D. B. Leake. Evaluating explanations. In *AAAI*, pages 251–255, 1988.

[22] D. B. Leake and C. C. Owens. Organizing memory for explanation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 8, 1986.

[23] O. Li, H. Liu, C. Chen, and C. Rudin. Deep learning for case-based reasoning through prototypes: a neural network that explains its predictions. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

[24] L. Malandri, F. Mercorio, M. Mezzanzanica, and N. Nobani. Convxai: a system for multimodal interaction with any black-box explainer. *Cognitive Computation*, 15(2):613–644, 2023.

[25] S. Mohseni, N. Zarei, and E. D. Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4):1–45, 2021.

[26] I. Nkisi-Orji, C. Palihawadana, N. Wiratunga, D. Corsar, and A. Wijekoon. Adapting semantic similarity methods for case-based reasoning in the cloud. In *Case-Based Reasoning Research and Development: 30th International Conference, ICCBR 2022, Nancy, France, September 12–15, 2022, Proceedings*, pages 125–139. Springer, 2022.

[27] I. Nkisi-Orji, C. Palihawadana, N. Wiratunga, A. Wijekoon, and D. Corsar. Failure-driven transformational case reuse of explanation strategies in cloodcbr. In *International Conference on Case-Based Reasoning*, pages 279–293. Springer, 2023.

[28] C. Nugent and P. Cunningham. A case-based explanation system for black-box systems. *Artificial Intelligence Review*, 24:163–178, 2005.

[29] E. Plaza and J.-L. Arcos. Constructive adaptation. In *European Conference on Case-Based Reasoning*, pages 306–320. Springer, 2002.

[30] R. C. Schank and D. B. Leake. Creativity and learning in a case-based explainer. *Artificial intelligence*, 40(1-3):353–385, 1989.

[31] R. C. Schank, A. Kass, and C. K. Riesbeck. *Inside case-based explanation*. Psychology Press, 2014.

[32] M. Schrepp, A. Hinderks, and J. Thomaschewski. Applying the user experience questionnaire (ueq) in different evaluation scenarios. In *Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience: Third International Conference, DUXU 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings, Part I 3*, pages 383–392. Springer, 2014.

[33] B. Smyth and M. T. Keane. A few good counterfactuals: generating interpretable, plausible and diverse counterfactual explanations. In *International Conference on Case-Based Reasoning*, pages 18–32. Springer, 2022.

[34] M. Van Den Haak, M. De Jong, and P. Jan Schellens. Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behaviour & information technology*, 22(5):339–351, 2003.

[35] A. Vats, A. Mohammed, M. Pedersen, and N. Wiratunga. This changes to that: Combining causal and non-causal explanations to generate disease progression in capsule endoscopy. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[36] A. Wijekoon, D. Corsar, and N. Wiratunga. Behaviour trees for creating conversational explanation experiences. *arXiv preprint arXiv:2211.06402*, 2022.

[37] A. Wijekoon, N. Wiratunga, K. Martin, D. Corsar, I. Nkisi-Orji, C. Palihawadana, D. Bridge, P. Pradeep, B. D. Agudo, and M. Caro-Martínez. Cbr driven interactive explainable ai. In *International Conference on Case-Based Reasoning*, pages 169–184. Springer, 2023.

[38] A. Wijekoon, D. Corsar, N. Wiratunga, K. Martin, and P. Salimi. Tell me more: Intent fulfilment framework for enhancing user experiences in conversational xai. *arXiv preprint arXiv:2405.10446*, 2024.

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

[40] N. Wiratunga, A. Wijekoon, I. Nkisi-Orji, K. Martin, C. Palihawadana, and D. Corsar. Discern: Discovering counterfactual explanations using relevance features from neighbourhoods. In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1466–1473. IEEE, 2021.