Conceptual Modelling of Explanation Experiences Through the iSeeOnto Ontology

Marta Caro-Martínez[1,*], Anjana Wijekoon[2], Juan A. Recio-García[1], David Corsar[2] and Ikechukwu Nkisi-Orji[2]

[1]*Department of Software Engineering and Artificial Intelligence, Instituto de Tecnologías del Conocimiento, Universidad Complutense de Madrid, Spain*

[2]*School of Computing, Robert Gordon University, Aberdeen, Scotland*

Abstract
Explainable Artificial Intelligence is a big research field required in many situations where we need to understand Artificial Intelligence behaviour. However, each explanation need is unique which makes it difficult to apply explanation techniques and solutions that are already implemented when faced with a new problem. Therefore, the task to implement an explanation system can be very challenging because we need to take the AI model into account, user's needs and goals, available data, suitable explainers, etc. In this work, we propose a formal model to define and orchestrate all the elements involved in an explanation system, and make a novel contribution regarding the formalisation of this model as the iSeeOnto ontology. This ontology not only enables the conceptualisation of a wide range of explanation systems, but also supports the application of Case-Based Reasoning as a knowledge transfer approach that reuses previous explanation experiences from unrelated domains. To demonstrate the suitability of the proposed model, we present an exhaustive validation by classifying reference explanation systems found in the literature into the iSeeOnto ontology.

1. Introduction

XAI or eXplainable Artificial Intelligence is one of the most remarkable fields in computer science and Artificial Intelligence (AI) due to its application in many critical domains such as medicine, defence or industry [**?**]. Understanding the reason behind the behaviour of an AI can make a difference in the effectiveness of that model. A central challenge to provisioning XAI is the variety of situations where we need explanations, regarding different domains, users, goals or problems to solve [1]. This exacerbates the decision making task of selecting the right explanation approaches to apply. The *iSee* (Intelligent Sharing of Explanation Experience by Users for Users) project[1] was proposed with the objective of tackling this challenge. The final goal of the iSee project is to build a platform where users can share and reuse *explanation*

✉ martcaro@ucm.es (M. Caro-Martínez); a.wijekoon1@rgu.ac.uk (A. Wijekoon); jareciog@ucm.es (J. A. Recio-García); d.corsar1@rgu.ac.uk (D. Corsar); i.nkisi-orji@rgu.ac.uk (I. Nkisi-Orji)

[1]http://isee4xai.com

experiences, a concept that comprises all the elements involved in the explanation process: input data, user's goals, human-computer interaction, AI model, explanation algorithms, explanation performance, etc. Here, Case-Based Reasoning (CBR) appears as a natural approach to reuse explanation experiences developed by other users for different models and domains.

To support this CBR process, one of the main results we want to achieve in iSee is the formalisation of a conceptual model that defines an explanation experience. This model has been instantiated into the iSeeOnto ontology, that is able to represent the different conceptual dimensions that describe an explanation experience, mainly: (1) the AI model to explain; (2) the explainer we need to do that and; (3) the user who needs the explanation. Moreover, for each explanation experience, iSeeOnto defines the solution of the problem we want to solve, i.e. the features of the explainer that solves the problem, its implementation, and the methodology to evaluate its performance when deployed into an actual use case (the result of the experience). This way, iSeeOnto provides the conceptual support to describe the explanation experiences as cases, enabling the application of the CBR paradigm to reuse them.

The paper is structured as following. Section 2 describes some literature related to XAI and explanation conceptualisation. Section 3 depicts the current version of the iSeeOnto and the main concepts defined in the ontology. In Section 4, we validate our model, describing explanation systems proposed in the state-of-the-art with our ontology, and building a first approach of the case base we need to build for our iSee platform. Finally, we extract some conclusions from the work done in Section 5.

2. Background

There is an appreciable amount of surveys, reviews and taxonomies about discovering the concepts related to XAI methods. A recent survey explores XAI methods for deep learners which proposes a taxonomy of three main concepts: *scope* which identifies the entity explained by the method; *methodology* which recognises the algorithmic approach to explanation; and *usage* which refers to the applicability of the method to any or specific methods [2]. In addition to characterising XAI methods, authors of [3] also recognise user needs (i.e. intent) such as effectiveness, transparency, persuasiveness and usefulness and the type of explanations that satisfy these needs. The match between intents and explanations were derived from literature that used either user studies or empirical evaluations. Conceptualisation of XAI by the authors of [4] is another prominent contribution in XAI. It depicts explanations in machine learning and deep learning, while classifying XAI methods with respect to several aspects, such as explanation scope, transparency, data, and visualisation modes, etc.

Our previous work [5], also developed a conceptual model, which extended previous work [3, 4] to include new concepts to capture user knowledge. In ExRecOnto ontology we guide the design of explanation systems for recommenders through an ontology-based methodology. When developing iSeeOnto, we further extend our work [5] to capture facets of XAI as a user experience. We also refer to different aspects highlighted in XAI taxonomies found in literature such as user requirements [6] and evaluation of XAI methods [7].

One of the goals of iSee is to develop a CBR system that recommends appropriate XAI methods for AI systems by capturing successful cases of building XAI systems (i.e. explanation
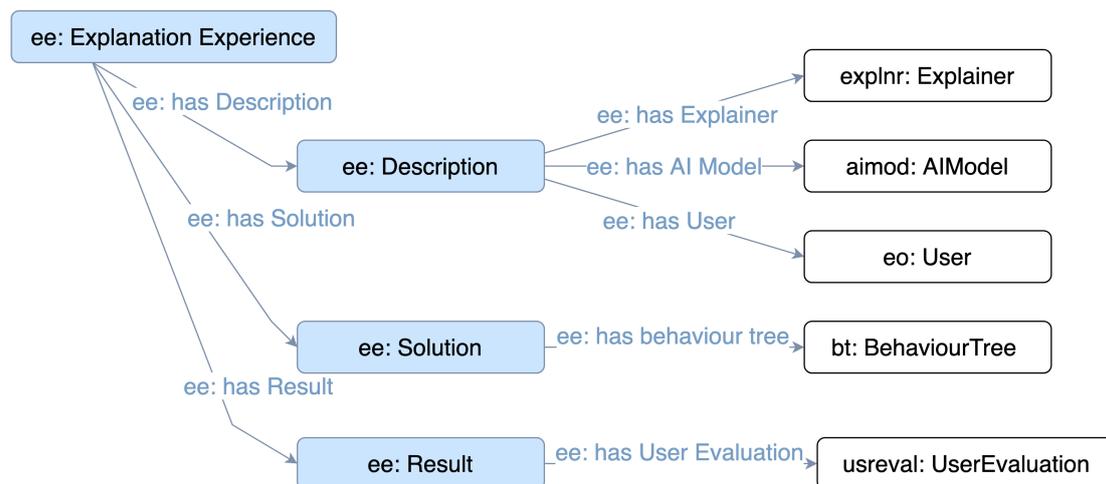
Figure 1: Explanation Experience Ontology

experience). CBR systems which make use of ontologies is a common approach seen in literature. We view the usage of iSeeOnto in two main knowledge sources for the CBR system [8]. Firstly, an ontology can be used as the vocabulary that represents the cases, in a similar way to the work in [9] and [10]. Secondly, iSeeOnto can be viewed as similarity knowledge for retrieval and reuse similar to work in [8] and [11]. In this paper we focus on the first knowledge container and we explore how to represent an XAI user experience as a case using iSeeOnto.

3. iSeeOnto

iSeeOnto is an ontology that describes explanation experiences for user-centered XAI. It enables the definition of explanation experiences as cases, consisting of description, solution and result, that can be reused by the iSee CBR engine.

Figure 1 presents the complete view of the iSeeOnto and how it forms the case structure for iSee CBR system. A case describes an explanation experience that captures how an explanation strategy satisfies certain explanation needs. Thus, the ontology defines three main features that describe an explanation experience: (1) the *AI model* we need to explain; (2) the *explainer* features we require to explain the AI model and; (3) the *user* and their explanation requirements.

The solution of an explanation experience formalises how we compose several explanation artefacts to implement a working system that achieves the requirements defined in the description. This way, solutions must capture the combination of explanation components, user evaluation methods, and user-explainer interaction. Workflows are the most common paradigm to implement such needs as they define the specific sequence of steps (or tasks) involved in the process of getting a work done. However, workflows put the user's decision at the centre, not the process itself, and are typically coupled to a particular application instead of being reusable as required by the iSee platform. Instead, we propose the use of Behaviour Trees (BTs) that
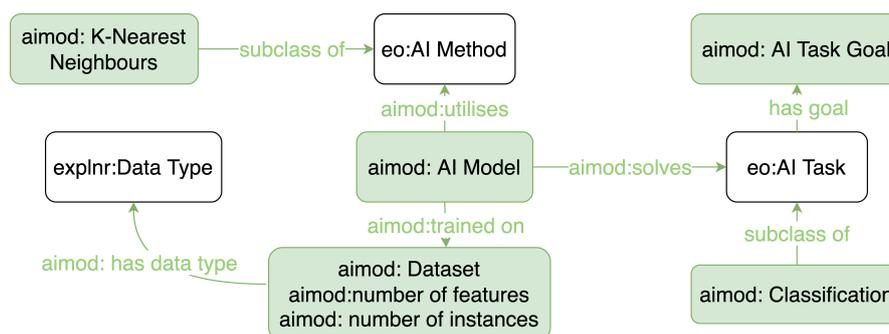
Figure 2: AI Model Ontology

are mathematical models of plan execution in a modular fashion. They define how a given objective should be achieved in a domain independent way, and in how many different states the information can be during that process. As a result, the process that is described using BTs is independent of the end-user's roles and is modular and reusable, the latter being a major requirement of the CBR adaptation process.

Finally, the *UserEvaluation* describes how an *Explanation* can be evaluated, defining the *Metric* and the *Dimension*. Next, following subsections describe each one of the previous concepts.

3.1. AI Model

AIModel concept and related features are presented in Figure 2. *AIModel* utilises an *AIMethod* which has sub classes like k-NN, SVM and Neural Networks. *AIModel* is trained on a *Dataset* which has a *DataType* (i.e. tabular, image, text, etc.) and characterised by *number of features* and *number of instances*. *AIModel* solves an *AITask* such as classification, regression or anomaly detection which is also associated with an *AITaskGoal*. iSeeOnto also characterises how an *AIModel* is evaluated (not included in Figure 2). The *AIModel* has *AIModelAssessmentResult* from when it is evaluated for a *AIModelAssessmentDimension* such as performance, bias or robustness using a *AIModelAssessmentMetric* like accuracy, f1-score or recall.

3.2. Explainer

Considering the *Explainer*, iSeeOnto defines several concepts related to its features as shown in Figure 3. The *Explainer* utilises an *ExplainabilityTechnique* that generates an *Explanation*. An *Explanation* can be a Feature Importance Explanation, a Contextual Explanation or a Instance-based Explanation, etc. An *ExplainabilityTechnique* can be applicable only to a certain AI method or an AI task. The *ExplainabilityTechnique* generates an *Explainer* which has a representation characterised *InformationContentEntity*. A few sub classes of *InformationContentEntity* are visual, text, annotations and charts.

The *ExplainabilityTechnique* is characterised by many features. *Portability* indicates if the technique is AI model-agnostic (works for any AI model), AI model-specific (only works for a specific AI model) or model-class specific (only works for a specific type of AI models). *ExplainerConcurrentness* is ante-hoc when the explanation is generated by the AI model itself; it
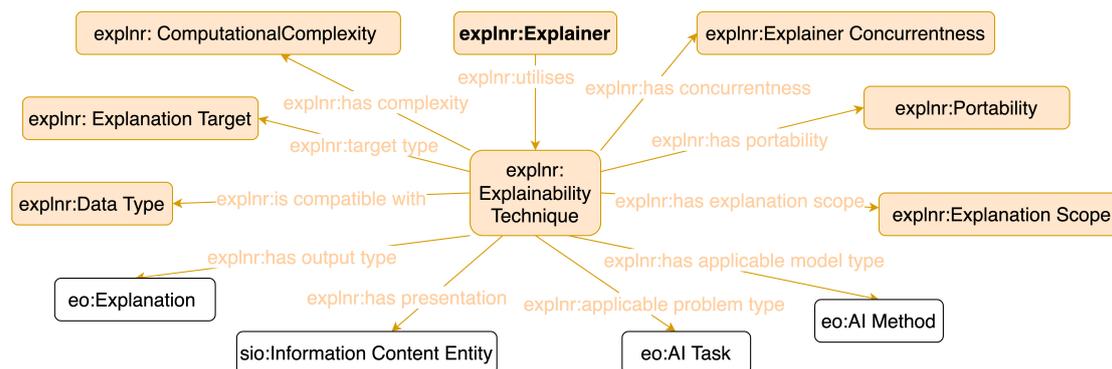
Figure 3: Explainer Ontology

is considered post-hoc when the explainability technique is independent of the AI model. The *ExplainabilityTechnique* has an *ExplanationScope*: local when it explains a single data point or a single prediction; global when it explains the behaviour of the AI model or data as a whole; and cohort when the explanation is related to a subset of predictions or data. An *ExplanationTarget* is related to the previous concept; it determines the target of the explanation, the predictions, the model or the data.

3.3. User

Regarding the user, iSeeOnto considers concepts depicted in Figure 4. A user has an *Intent* that is expressed as a *UserQuestion*. An *Intent* describes the need for an explanation which can fall under a category such as Transparency, Trust, Effectiveness, Efficiency, Scrutability, Satisfaction, Persuasiveness, Education or Debugging. The *UserQuestion* can take the form of a How, Why, What, When, Where, or What-if question. The *UserQuestionTarget* of the *UserQuestion* points to the aspect of the AI model's behaviour is being questioned by the user (i.e. system recommendation, AI model or data. iSeeOnto also recognises that an *Explanation* addresses a certain *UserQuestion*.

Knowledge possessed by the *User* can take two forms: Domain Knowledge and AI Knowledge both of which can be measured as *low*, *medium* or *high*. The *User* has some resources identified by *TechnicalFacilities* such as touch, audio or visual determine which explanation modalities can be presented. For instance, an *Explanation* that has a interactive presentation is only suitable if the user has a touch or clickable interface.

3.4. Behaviour Tree

Figure 5 captures the different Nodes and properties of a *BehaviourTree* that is required to represent an explanation strategy. A BT can have multiple *Tree*s, each consists of multiple *Node*s. There are four main types of Nodes: *CompositeNode* and *DecoratorNode* for controlling flow between nodes; *ActionNode* for performing an action such as retrieving and displaying an
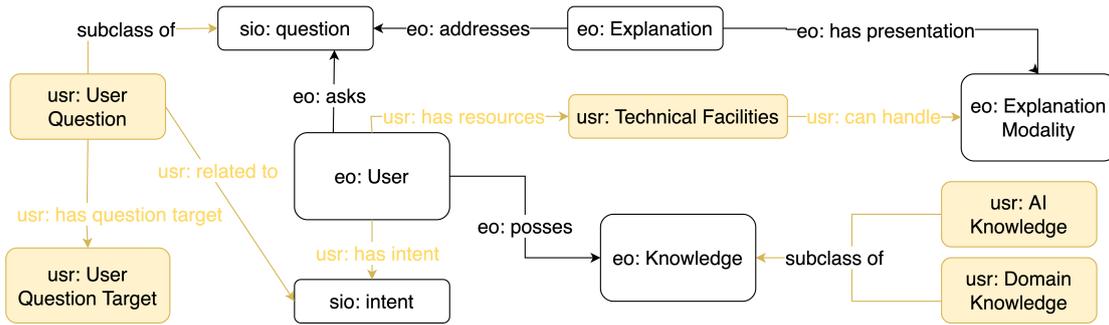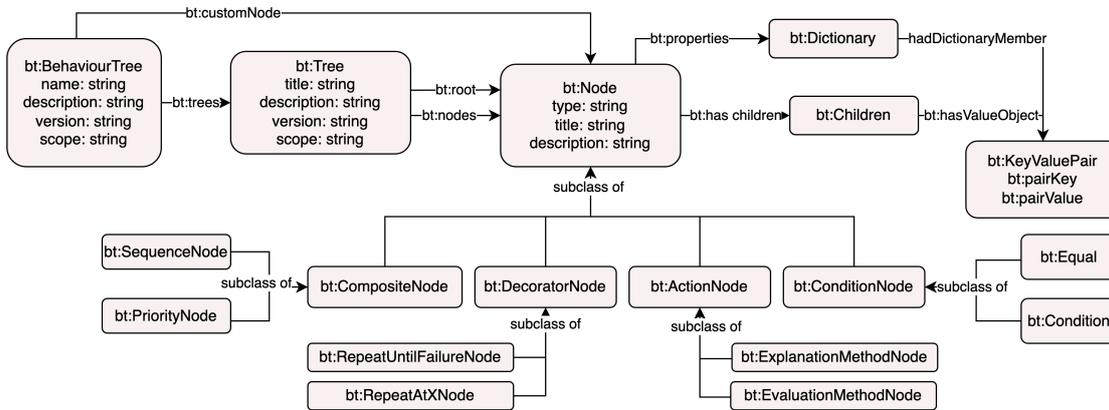
Figure 4: User Ontology

Figure 5: Behaviour Tree Ontology

explanation or executing an evaluation questionnaire; and *ConditionNodes* for controlling the entry to a sub-tree.

Sub classes of *CompositeNode* and *DecoratorNode* concepts carry the standard functionality as intended in generic Behaviour trees. In this paper we highlight few custom nodes for Explanation Strategy representation. The *ExplanationMethodNode* represents an action node that invokes an API call to a URL (in node property *URL*) that will fetch an explanation given pre-requisite parameters (also included in node properties). The *EvaluationMethodNode* will fetch a questionnaire from a URL that will be executed as a sub-tree. The *ConditionNode* will control the entry to a sub-tree which is also known as memory controlled flow for efficient execution of BTs.

3.5. User Evaluation

The *UserEvaluation* describes how an *Explanation* is evaluated, defining the *Metric* and the *Dimension.* An *Explanation* is annotated by a *UserEvaluationResult* which is based on a *Metric.*
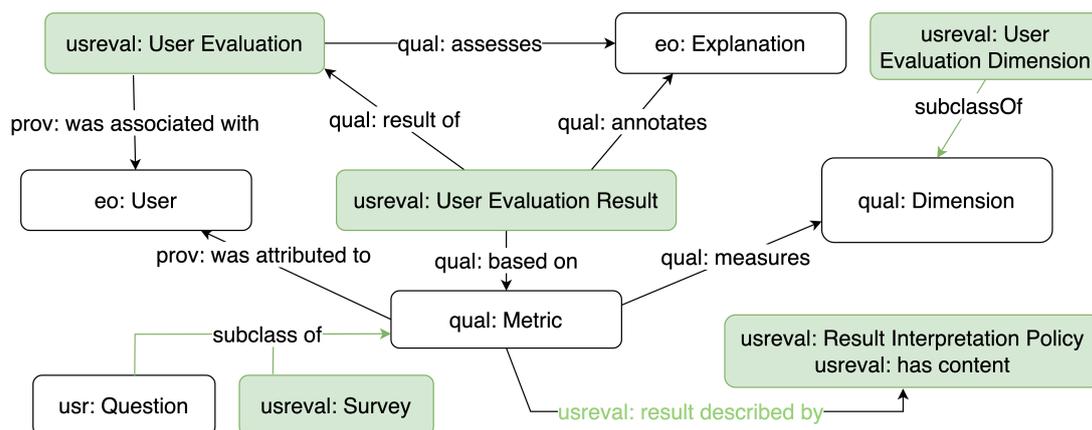
Figure 6: User Evaluation Ontology

The *Metric* can be either a *Question* or *Survey* which consists of multiple questions. The *Metric* is designed to measure a *Dimension* such as Satisfaction, Goodness or Trust. The *UserEvaluationResult*s are collated using a *ResultInterpretationPolicy* which determines the overall success of the *Explanation*, and subsequently the suitability of the *Explainer* for the associated *User*.

4. Validation

To validate the completeness of iSeeOnto regarding its capability to represent an explanation experience properly, we have carried out a validation consisting of the description of thirteen explanation approaches proposed in the literature with the concepts defined in iSeeOnto. We selected these explanation approaches trying to cover a wide variety of features and types. We extracted most of them from the work by Barredo-Arrieta et al. [4]. We also selected some works we used in the validation carried out in our previous paper [5] for the recommender systems domain. Finally, a few of the publications were selected making a systematic search on Google Scholar from December 2021 until February 2022. The keywords used in the search were: "XAI taxonomy", "explanation systems artificial intelligence" and "explanations artificial intelligence" combining these concepts with the ones defined in the taxonomy by Barredo-Arrieta et al. We filtered results for publications not older than ten years. This way, we could represent an amount of different explanation approaches with our ontology, validating its completeness.

Following, we describe every explanation approach that was successfully represented using iSeeOnto and the concepts specified in Section 3.

Ontology					
explanation Experience.rdf	**ExplanationExperience**				
	hasDescription	**Description**			
aimodel.rdf		hasAIModel	**AIModel**		
			trained on	**Dataset**	
				hasDataType	**DataType**
					number of features
					number of instances
			solves	**AITask**	
				hasType	**AITaskType**
				hasGoal	**AITaskGoal**
			utilises	**AIMethod**	
				hasType	**AIMethodType**
aimodelevaluation.rdf		annotated by	AIModelAssessmentResult		
			basedOn		**AIModelAssessmentMetric**
			measures		**AIModelAssessmentDimension**
explainer.rdf		needExplainer	**Explainer**		
				hasOutputType	***Explanation***
				hasPortability	**Portability**
				hasConcurrentness	**ExplainerConcurrentness**
				hasPresentation	**InformationContentEntity**
				hasExplanationScope	***Explanation Scope***
				targetType	**Explanation Target**
user.rdf		hasUser	**User**		
			asks	**UserQuestion**	
				hasTarget	**UserQuestionTarget**
				hasType	**QuestionType**
			has intent	**Intent**	
				hasType	IntentType
			has resources	**Technical Facilities**	
				can handle	**ExplanationModality**
			possess	**Domain Knowledge**	
				level of domain knowledge	**KnowledgeLevel**
				level of AI knowledge	**KnowledgeLevel**
behaviour_tree.rdf	hasSolution	**Solution**			
		hasExplainer	**Explainer**		
			utilises	**ExplainabilityTechnique**	
				hasType	**ExplainabilityTechniqueType**
				hasOutputType	**Explanation**
				hasPortability	**Portability**
				hasConcurrentness	**ExplainerConcurrentness**
				hasPresentation	**InformationContentEntity**
				hasExplanationScope	***Explanation Scope***
				targetType	**Explanation Target**
				isCompatiblewith	
				FeatureTypes	**DataType**
				hasComplexity	**ComputationalComplexity**
userevaluation.rdf	hasDone	**UserEvaluation**			
		basedOn	**Metric**		
			measures	**Dimensions**	

Figure 7: Schema of the concepts used to validate iSeeOnto

The details of this classification cannot be included in this paper due to space restrictions but are available on GitHub[2]. The schema of the concepts used to classify each explanation approach are presented in Figure 7. Below we introduce the explainers we have examined.

PSIE [12]. This paper proposes to recommend movies to groups based on social knowledge.

DisCERN [13]. The goal of this paper is to predict lung cancer risk given clinical data of patients.

BTTelecom [14]. In this paper, authors propose to recommend engineering notes to desk support staff to help on-site engineers.

SciNet [15]. In this work, the objective is to determine the most related documents given a set of keywords. One of the most important features of this approach is that users can interact with the interface to manipulate the keywords and change the search result.

TalkExplorer [16]. This paper proposes to recommend scientific publications based on content and social connections. In this case, an explanation system is incorporated into a conference recommender system. The conferences are grouped in bubbles according to their content. Users can interact with the interface to get more details of the recommendation.

IntGradImage, **IntGradRetinopathy** and **IntGradTextClassification** [17]. In this work, the authors propose the explanation technique Integrated Gradients. With this technique, they are able to predict the category of a given image, predict if a given medical image contains diabetic retinopathy and predict the question category based on question text, respectively for each approach.

KimEtAlMethod [18]. The system proposed in this paper makes acceleration or change course decisions in a self-driving car based on video.

iBCM [19]. The main goal of this work is to cluster student assignment submissions to design grading rubric or to compose feedback.

InNoCBR [20]. This publication describes a CBR system with explanations for detecting healthcare associated infections (HCAIs). The system predicts patient's infection based on a clinical, laboratory, and medico administrative based data.

DeepSHAPGlobal and **DeepSHAPLocal** [21]. In this work, the SHAP method is proposed to predict patient mortality based on clinical, nutritional and behavioural factors.

From this representation of such a heterogeneous collection of explainers in iSeeOnto, we can support its semantic completeness and extract some conclusions about the features of the explanation systems we have been validated.

Regarding the *AIModel* and the *AITask* concepts, we can see that almost half of the approaches explain AI models related to disease prediction using classification tasks. This is also the most common *AITaskType*: multi-class and binary classification.

---

[2]https://github.com/isee4xai/iSeeOnto/tree/main/case-structure

Considering the *Explainer* features, there are several types of explanations, not having a type of *Explanation* that stands out over others. However, we can observe that some of the explaination approaches for recommender systems are neighbourhood-based explanations. Regarding *Portability*, we have more post-hoc systems than ante-hoc systems. When the explainer is ante-hoc, the explanation is generated by the AI model itself, so the explainer cannot be decoupled, and therefore it not reusable as the solution for a different problem by a CBR system. According to the *InformationContentEntity* concept, *any entity* is the widest option although specific values such as *text* and *image* are other options that are also quite common. The *ExplanationScope* is one of the concepts where we can see uniform results: most of the explainers are local. The same happens with *ExplanationTarget*, where we can see that most of the explainers has prediction as target.

Taking into account the classification with concepts related to the *User*, we can determine we have been able to represent a broad variety of users, depending on the domain of application. However, the majority of the *UserQuestion* of the approaches are *Why* questions, having only two *What/How* questions, one *Why/How* question and one *How/What-if* question. Then, we can conclude that one of the main users' goals seems to be the understanding of the underlying *AIModel*. One of the most repeated *Intents* of the *User* is Transparency, followed by Trust, which makes sense considering the type of questions the users want to answer about the AI model in our validation. The *TechnicalFacilities* required by the user are unanimous: in all the explanation experiences studied, users need *Screen Display* to watch the explanation. Regarding the user knowledge level, most of the users have high level of domain knowledge, and low knowledge level in AI.

Considering the *Solution* concept, we have to analyse the results from the *UserEvaluation*. As we can see in the spreadsheet, half of the explanation experiences studied have not carried out an evaluation. There are five approaches evaluated, all of them use questionnaires as the evaluation metric. The dimensions measured are heterogeneous, being *Usefulness* and *Efficiency* two of the most common dimensions among the explanation experiences evaluated. Finally, we have to discuss the code availability of the implementation of the solution: only seven of the explanation experiences have the source code available.

5. Conclusions

Nowadays, trust in AI systems is essential for the acceptance of their predictions by the final user. Therefore, including explanations in intelligent systems is critical to make this type of systems effective.

Fortunately, the XAI community is generating novel explanation methods continuously that address this problem. However, it raises the challenge to find the most suitable explanation strategy for each AI system, being CBR a natural solution for reusing previous explanation experiences. Developing such solution is the goal of the *iSee project*. As the first step, we have developed the iSeeOnto ontology, which formalises a conceptual model to define and describe explanation strategies and all the elements involved: the AI model to explain, the explainer features and the user requirements.

To check the completeness of iSeeOnto, it has been validated by describing several repre-

sentative explanation approaches extracted from the state-of-the-art. The secondary goal of this validation is the collection of an initial case base for the future iSee's CBR platform. From the results of this validation, we can conclude that iSeeOnto is able to classify all of these approaches. Therefore we can propose our ontology as a guideline for designing explanation strategies through an ontology-based methodology.

Finally, as future work, we want to extend the validation and the resulting case base with further use cases. In addition, we want to establish the retrieval strategy required to reuse these explanation experiences for different domains, AI models and user's intends.

Acknowledgments

References

[1] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): Toward medical XAI, IEEE transactions on neural networks and learning systems 32 (2020) 4793–4813.

[2] A. Das, P. Rad, Opportunities and challenges in explainable artificial intelligence (XAI): A survey, arXiv preprint arXiv:2006.11371 (2020).

[3] I. Nunes, D. Jannach, A systematic review and taxonomy of explanations in decision support and recommender systems, User Modeling and User-Adapted Interaction 27 (2017) 393–444.

[4] A. Barredo-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, Information fusion 58 (2020) 82–115.

[5] 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, Journal of Artificial Intelligence Research 71 (2021) 557–589.

[6] S. Chari, O. Seneviratne, D. M. Gruen, M. A. Foreman, A. K. Das, D. L. McGuinness, Explanation ontology: a model of explanations for user-centered AI, in: International Semantic Web Conference, Springer, 2020, pp. 228–243.

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

[8] J. A. Recio-Garía, B. Díaz-Agudo, Ontology based CBR with jCOLIBRI, in: International Conference on Innovative Techniques and Applications of Artificial Intelligence, Springer, 2006, pp. 149–162.

[9] Y. Guo, J. Hu, Y. Peng, A CBR system for injection mould design based on ontology: a case study, Computer-Aided Design 44 (2012) 496–508.

[10] A. Bouhana, A. Zidi, A. Fekih, H. Chabchoub, M. Abed, An ontology-based CBR approach

for personalized itinerary search systems for sustainable urban freight transport, Expert systems with applications 42 (2015) 3724–3741.

[11] D. Wang, Y. Xiang, G. Zou, B. Zhang, Research on ontology-based case indexing in CBR, in: 2009 international conference on artificial intelligence and computational intelligence, volume 4, IEEE, 2009, pp. 238–241.

[12] L. Quijano-Sanchez, C. Sauer, J. A. Recio-Garcia, B. Diaz-Agudo, Make it personal: a social explanation system applied to group recommendations, Expert Systems with Applications 76 (2017) 36–48.

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

[14] K. Martin, A. Liret, N. Wiratunga, G. Owusu, M. Kern, Evaluating explainability methods intended for multiple stakeholders, KI-Künstliche intelligenz 35 (2021) 397–411.

[15] D. Glowacka, T. Ruotsalo, K. Konuyshkova, K. Athukorala, S. Kaski, G. Jacucci, Directing exploratory search: Reinforcement learning from user interactions with keywords, in: Proceedings of the 2013 international conference on Intelligent user interfaces, 2013, pp. 117–128.

[16] K. Verbert, D. Parra, P. Brusilovsky, E. Duval, Visualizing recommendations to support exploration, transparency and controllability, in: Proceedings of the 2013 international conference on Intelligent user interfaces, 2013, pp. 351–362.

[17] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International conference on machine learning, PMLR, 2017, pp. 3319–3328.

[18] J. Kim, A. Rohrbach, T. Darrell, J. Canny, Z. Akata, Textual explanations for self-driving vehicles, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 563–578.

[19] B. Kim, E. Glassman, B. Johnson, J. Shah, ibcm: Interactive bayesian case model empowering humans via intuitive interaction (2015).

[20] H. Gómez-Vallejo, B. Uriel-Latorre, M. Sande-Meijide, B. Villamarín-Bello, R. Pavon, F. Fdez-Riverola, D. Glez-Pena, A case-based reasoning system for aiding detection and classification of nosocomial infections, Decision Support Systems 84 (2016) 104–116.

[21] H. Chen, S. Lundberg, S.-I. Lee, Explaining models by propagating shapley values of local components, in: Explainable AI in Healthcare and Medicine, Springer, 2021, pp. 261–270.