Towards FAIR Explainable AI: a standardized ontology for mapping XAI solutions to use cases, explanations, and AI systems

Ajaya Adhikari
TNO
The Hague, The Netherlands
ajaya.adhikari@tno.nl

Edwin Wenink
TNO
The Hague, The Netherlands
edwin.wenink@tno.nl

Jasper van der Waa
TNO
The Hague, The Netherlands
jasper.vanderwaa@tno.nl

Cornelis Bouter
TNO
The Hague, The Netherlands
cornelis.bouter@tno.nl

Ioannis Tolios
TNO
The Hague, The Netherlands
ioannis.tolios@tno.nl

Stephan Raaijmakers
TNO
The Hague, The Netherlands
stephan.raaijmakers@tno.nl

ABSTRACT

Several useful taxonomies have been published that survey the eXplainable AI (XAI) research field. However, these taxonomies typically do not show the relation between XAI solutions and several use case aspects, such as the explanation goal or the task context. In order to better connect the field of XAI research with concrete use cases and user needs, we designed the ASCENT (Ai System use Case Explanation oNTology) framework, which is a new ontology and corresponding metadata standard with three complementary modules for different aspects of an XAI solution: one for aspects of AI systems, another for use case aspects, and yet another for explanation properties. The descriptions of XAI solutions in this framework include whether the XAI solution has a positive, negative, inconclusive or unresearched relation with use case elements. Descriptions in ASCENT thus emphasize the (user) evaluation of XAI solutions in order to support finding validated practices for application in industry, as well as being helpful for identifying research gaps. Describing XAI solutions according to the proposed common metadata standard is an important step towards the FAIR (Findable, Accessible, Interoperable, Reusable) usage of XAI solutions.

CCS CONCEPTS

• **Information systems** → **Ontologies**; *Web Ontology Language (OWL)*; • **Computing methodologies** → *Ontology engineering*; • **Human-centered computing** → Human computer interaction (HCI).

KEYWORDS

XAI ontology, FAIR, user-centered, ASCENT

PETRA '22, June 29-July 1, 2022, Corfu, Greece

ACM ISBN 978-1-4503-9631-8/22/06…$15.00
https://doi.org/10.1145/3529190.3535693

ACM Reference Format:
Ajaya Adhikari, Edwin Wenink, Jasper van der Waa, Cornelis Bouter, Ioannis Tolios, and Stephan Raaijmakers. 2022. Towards FAIR Explainable AI: a standardized ontology for mapping XAI solutions to use cases, explanations, and AI systems. In *The15th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '22), June 29-July 1, 2022, Corfu, Greece.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3529190.3535693

1 INTRODUCTION

The field of eXplainable Artificial Intelligence (XAI) is rapidly progressing towards a large variety of technologies that generate human-interpretable explanations from AI systems. This momentum is boosted by the establishment of national and international regulations and legal frameworks that stress the importance of *trustworthy and responsible* AI. An example is the recent draft of the AI Act by the European Commission [10] which states that AI should be developed and applied in a responsible and trustworthy manner, generally following the moral imperative that AI should "do good" [9, 23]. Transparency and explainability are seen as two critical ingredients for this mission. XAI technologies are receiving increasingly more attention from the industry and academia, and many implementations are readily available (see Rothman [20] for an overview). Given the vast amount of XAI technologies, multiple survey papers [2, 3, 13] aim to provide an overview using different taxonomies. Most of these taxonomies, however, categorize XAI solutions mainly based on model characteristics and explanation types with little focus on the end-user. Moreover, they typically do not show the relation between XAI solutions and several use case aspects, such as the explanation goal or the task context. As a result, many of these taxonomies - viewed in isolation - mainly serve data scientists. In order to better connect the field of XAI research with use cases in industry, we argue that it is beneficial to also explicitly model the implications of an XAI solution for different use case aspects according to user evaluations, alongside categorizations of the AI model and the explanation that is desired. Because there are many different configurations of use cases, AI models, and explanation algorithms, we propose to model XAI solutions as instances of an ontology.

In this work, we propose the ASCENT (Ai System use Case Explanation oNTology) framework which consists of an ontology and

corresponding metadata standard for annotating XAI solutions according to three ontology modules, namely *AI System*, *Use Case*, and *Explanation Algorithm*. This framework suits the growing focus within the XAI field towards more rigorous evaluation [8], incorporating scientific theories [12, 15, 26] and a human-centered approach [9, 22]. The AI System module aligns with the technology-centered approach within XAI, as it describes the properties that signal what explanation can be generated (e.g. the type of data and model used). The Use Case module aligns with the growing human-centered approach to XAI and describes the important use case properties, which should be taken into account when searching for suitable XAI methods (e.g. the goal of an explanation). Finally, the Explanation Algorithm module describes the aspects of an explanation that need to be considered given these use case and AI system properties (e.g. the explanation generating method and the explanation's modality).

The ASCENT framework allows a more FAIR [30] usage of XAI solutions by providing rich machine-readable metadata for XAI solutions (Findable), online access to our metadata standard (Accessible), the ontology in a commonly used OWL standard (Interoperable), and extension possibilities (Reusable). This paves the way for the creation of a communal repository of XAI solutions described according to a common metadata standard that is relevant for industry applications.

This paper is structured as follows. In the next section we address relevant literature on taxonomies and categorizations of the XAI field and emphasize our contribution. In Section 3, we present and motivate the proposed ontology and its modules. Section 4 elaborates on how the ASCENT framework is used in creating metadata of XAI methods. Finally, in Section 5, we will discuss and conclude on the proposed methodology.

2 BACKGROUND

Some recent surveys of XAI methods use taxonomies to provide a clear overview of the XAI research field. Arrieta et al. [2] provide an in-depth overview of existing XAI methods, accompanied by two taxonomies. The first taxonomy provides a broad overview that allows for traversing a path through the taxonomy tree, based on the type of XAI solution or the model type. At the leaves, this taxonomy presents methods and their paper references, with additional color coding for the data type the method is applicable to. Belle and Papantonis [3] offer a similar but more succinct taxonomy, where the scope is explicitly limited to explanations of AI techniques that rely on statistical association. A second taxonomy by Arrieta et al. [2] specifically limits the scope to XAI methods for Deep Learning (DL) models such as Deep Neural Networks (DNNs), because these models tend to be the most problematic in terms of interpretability. Jin et al. [14] develop a prototyping tool for explanations that uses an extended version of the taxonomy presented in [13]. This work defines prototypical forms of explanations with UI/UX design templates and associated XAI algorithms.

XAI methods are in general not plug-and-play. For instance, they may require consideration of various user roles with different explanation needs. We observe that existing taxonomies mainly support the XAI research community and data scientists, but do not sufficiently address the societal need for tools that aid in the responsible application of XAI by factoring in user and use case properties. The literature in which such taxonomies are presented *does* however typically recognise that explanation is a difficult concept studied not only in AI, but also in philosophy and social sciences [e.g. 15]. For example, Arrieta et al. [2] propose a definition of explainable AI in which the target audience is factored in. Belle and Papantonis [3] show awareness of users and use cases by discussing a use case of a hypothetical data scientist that has to solve questions posed by business managers. The prototyping process designed by Jin et al. [14] is user-focused and was tested in user studies, but at the cost of explicitly limiting its scope to XAI methods "that do not require technical knowledge to comprehend" [14]. Nevertheless, the taxonomies themselves do not support a mapping to use case properties and leave this to the discretion of the data scientist. We instead propose an ontology that explicitly models use case properties alongside explanation and AI system properties.

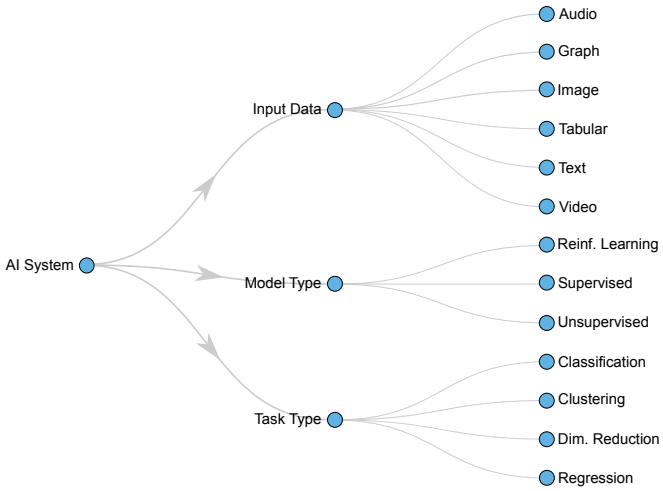
Explanation ontologies have been published by [24] and [6], with the former being reused in the latter. [24] is a general purpose ontology, whereas [6] is specifically designed for user-centered AI explanations and thus has a similar aim as ASCENT. Our ontology however focuses specifically on XAI solutions and, as a result, we include insights from the several existing XAI taxonomies in greater detail. Having this more specific scope may improve the practical usability of the ontology, as well as provide a more fine-grained XAI-specific variant compared to the more general explanation ontologies. Another contribution of our ontology is that relations between a given XAI solution and use case aspects indicate whether this relation is researched and, if so, validated. This way, user evaluations are an integral part in the description of XAI solutions, rather than an afterthought.

3 ONTOLOGY

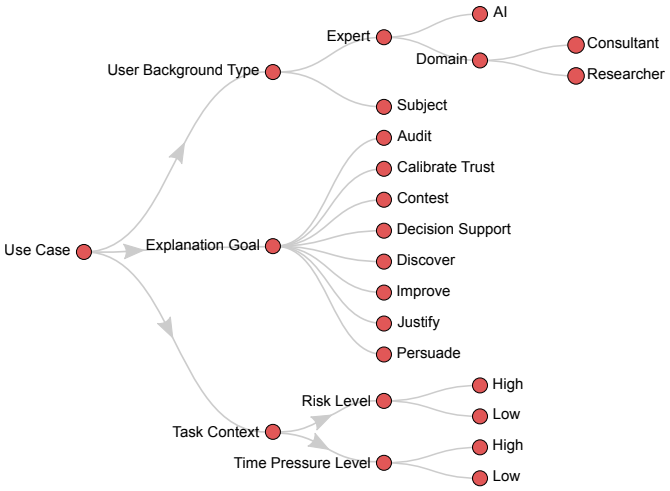
We define an overarching domain ontology for explanainable AI applications with three modules (or sub-ontologies), namely *Use Case* (figure 1b), *AI System* (figure 1a) and *Explanation Algorithm* (figure 1c).

The main goal of the ontology is to describe the collection of entities that play a role when an XAI solution has to be identified based on a broad spectrum of requirements. The AI system is the main concept, to which both the use case (including the goals of the users) and the explanation algorithms are related. Figure 1d illustrates the relations between the three modules. A Use Case *employs* an AI System, which *is employed in* that Use Case. An AI System *provides an explanation service* by offering an Explanation Algorithm that *explains* the AI System. Various AI systems may be (potentially) involved in a use case. Additionally, various explanation algorithms may be applicable on a single AI system.

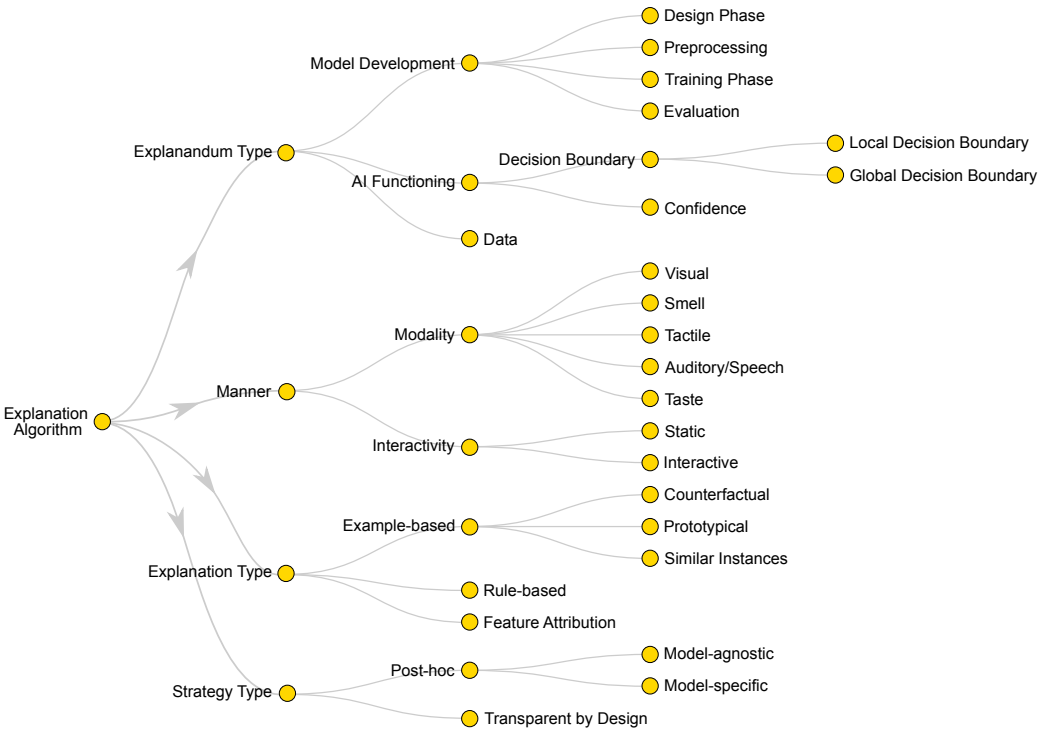
For each module we identify a first set of data elements that covers the majority of use cases, but we expect them to be expanded on for a specific use case. The main focus of our effort is not to replace existing taxonomies, but rather to incorporate their useful distinctions in a comprehensive framework that is useful for the development of practical applications. The AI system entity together with its associated use case and its associated explanation algorithms constitutes an environment that can be used in downstream applications. For example, it is possible to build a knowledge base



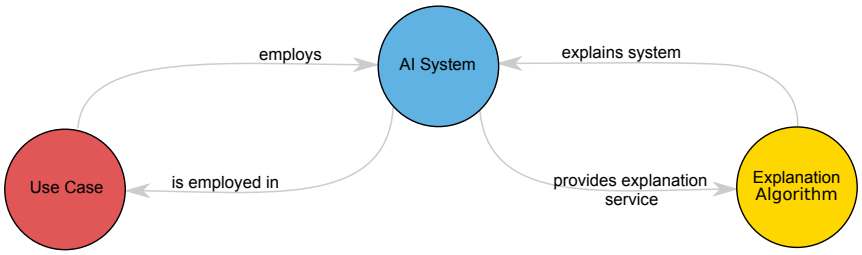
(a) The AI System module.



(b) The Use Case module.



(c) The Explanation module.



(d) Overarching ontology with three modules.

Figure 1: Visualisation of the ASCENT framework. In the module visualisations the connections without arrow indicate sub-class relations and connections with arrows property relations.

with a predetermined set of annotated explanation algorithms that we can query.

Note that even though we have displayed all data elements like a taxonomy for convenience, they are not used as such. A taxonomy is used for *classification*, which in turn assumes you end up at a single leaf node. However, in our figures multiple leaf nodes can be applicable simultaneously. For example, it is no redundancy that the word "data" appears both under *Explanandum Type* and in *Input Data*. The question which input data type is required by an AI system is independent from the question whether it is the goal of a particular explanation algorithm to provide insight into a data set. Also note that most links in the figures indicate subclass relations in the underlying ontology, but that links with arrows indicate property relations. See section 4 for more details on the ontology serialisation as OWL (Web Ontology Language).

3.1 Use Case

Three main use case characteristics are distinguished in the Use Case module, namely the *background* of the user, the *goal* of the explanation, and the *task context* (see figure 1b).

3.1.1 User Background. The background of the user affects which type of explanation is useful. We distinguish whether the user is an *expert* with respect to the application domain or the used AI, or the *subject* of the decisions supported by the AI, such as patients or consumers. For example, an explanation about the prediction of the most probable diagnosis cannot include complex medical terms when given to a patient, whereas a doctor probably needs a more in-depth and detailed explanation.

Within the expert group, an *AI* engineer typically focuses on debugging the AI model and how to improve its performance, while a *domain* expert wants to understand the underlying reasoning from the application perspective. Furthermore, *consultants* such as doctors and lawyers are typically interested in the reasoning behind the prediction of individual input samples, whereas *researchers* such as medical scientists and legislators are often also interested in more global explanations regarding multiple input samples.

3.1.2 Explanation Goal. An explanation serves a certain goal within a use case. We distinguish eight different explanation goals, which we illustrate with the use case of a doctor who is assisted by an AI model to diagnose and treat a patient:

Discover: AI can be used for the discovery of relevant patterns, such as clusters [7] and bias in the data [5]. For example, medical researchers may want to discover relevant relations between variables, such as diet and success of a treatment, by gaining insight from the AI model.

Decision Support: Explanations can aid decision making, for example by avoiding tunnel vision and providing actionable explanations. Doctors want to understand why a certain diagnosis is deemed most probable by the AI system, such that they can combine it with their own knowledge and expertise when deciding on the final diagnosis.

Calibrate trust: Explanations can aid in the appropriate use of an AI system by avoiding the extremes of over-reliance (blindly following the AI systems advice) and under-reliance (distrusting all output of the system) [27]. A doctor wants to understand when and why the AI system can be trusted.

Justify: After the doctor decided on the diagnosis with support of the AI system, an AI explanation can be used to justify the decision to the patient and in the medical report.

Audit: Auditors need to understand whether an AI system conforms to regulations. They for example may want to know which variables are used by the diagnosis model and how robust it is.

Improve: Explanations can provide insights to improve the underlying AI model. For example, if it turns out that the model makes more mistakes for a particular patient group, the developer may decide that more representative data is needed.

Contest: When contesting a decision, users want to understand the reasoning behind that decision. This goal is becoming more relevant with the increased use of AI systems in daily life and regulations such as the GDPR containing the right to contest. For example, if a patient does not agree with the diagnosis, the decision may be contested by inquiring into which features were deemed relevant by the AI model. A possible outcome may be that the patient finds out that some irrelevant, wrong or inappropriate features were used.

Persuade: The type of explanation supporting a certain recommendation can affect how persuasive it is [25]. For example, the doctor may want to persuade the patient to stop smoking when the patient is developing lung disease. This goal is especially prevalent in retail use cases for recommendation of goods or services [25].

3.1.3 Task Context. The context of the task for which an AI model is used has an effect on which types of explanation are useful. Firstly, in use cases with high *time pressure*, such as collaboration between an AI and emergency medical responders, there is little time to understand in detail why a certain action is suggested by the AI system, while a GP, on the other hand, typically has more time to understand the reasoning behind an AI suggested diagnosis.

In high *risk* domains such as health care, a detailed understanding of the underlying reasoning of an AI output is typically needed, as a mistake might have significant consequences. In low risk applications such as recommendation systems in retail, the consumer typically has less need for a detailed understanding, for example why a piece of clothing is recommended.

3.2 AI System

We define three main characteristics of an AI System which affect the possible XAI solutions that can be used to extract explanations, namely the *input data* of the model, the *model type* and the *task type* it performs.

3.2.1 Input data. Six types of input data are considered, namely *tabular*, *text*, *image*, *video*, *audio*, and *graph* data. The characteristics of the input data of the model influence the possibilities of explanations. For example, data types that are visual in nature, such as images and graphs, allow for more visual explanations such as an attention layer on top of an image [31] or a subgraph of nodes which the model considered important [32]. Conversely, tabular data with large amount of features can be challenging to visualize and require a different type of explanation.

3.2.2 Model Type. We distinguish *supervised*, *unsupervised* and *reinforcement* models, following the classification of Nicolas [18]. From an XAI perspective, supervised models have the advantage that their outputs are often human interpretable, even though the model itself can be very complex. Unsupervised models are more challenging and provide less interpretable output such assignments to opaque clusters, which in addition also should be presented in an interpretable way. At last, generating explanations for a Reinforcement Learning (RL) model can be challenging given the complex long-term behaviour of RL agents. For a more detailed classification, we refer to existing taxonomies which classify different types of AI models [e.g. 4, 18].

3.2.3 Task. When specified on a fine granularity, many hundreds machine learning (sub)tasks may be formulated. We instead focus on common machine learning tasks that are typically addressed either in the XAI literature or in industry. The most common tasks are *classification* and *regression* in a supervised setting, and *clustering* and *dimensionality reduction* in an unsupervised setting. Our ontology is easily extendable with more specific tasks.

3.3 Explanation Algorithm

We consider an explanation algorithm from the viewpoint of what is being explained (*explanandum*), the different *explanation types*, the various ways to present to the explainee (*manner*), and the *strategy type* used for explanation extraction.

3.3.1 Explanandum. Data: Because most AI applications nowadays are data-driven, it is important to explain properties of training and test data, e.g. for assessing potentially problematic bias in the AI system. Explanations may provide relevant data statistics as well as insights on how the data was collected. These insights may affect for which use cases a trained model is appropriate and can be recorded in standardized documentation of data sets [11].

Model Development: To ensure accountability for AI decisions it is also important to be able to explain and justify the design history of an AI system. We roughly distinguish four phases. In a *design phase* one may need to explain the choice for data set, whether a risk assessment was made, or why a particular model is appropriate given a concrete application domain. One may also ask how appropriate features were created and how possible issues are addressed during *preprocessing*. In the *training phase*, choices for hyper-parameter tuning or regularization are made. Although these are more technical in character, it is still important to be able to explain them e.g. explaining how undesirable consequences of overfitting on a particular subpopulation were mitigated. Finally, being able to explain the *evaluation* of the model is important for instilling trust in its responsible application. For example, why was the chosen evaluation metric appropriate given the goals and usage context of the AI? How is the error calibration of the model, e.g. was a model very certain of predictions that were in fact wrong?

AI functioning: A main focus of the XAI community has been to explain how an AI model functions, because contemporary subsymbolic AI systems often have millions of interacting parameters contributing to the final output. We are often also interested in explaining the model's *confidence* in its decisions for the sake of trust calibration. There is great variety in how models are explained, but in terms of what XAI methods explain, we distinguish *local* and *global decision boundaries*. Methods that explain decision boundaries locally aim to explain model behavior in a limited region of the model input space (e.g. the feature space close to a particular data instance), whereas explanations on global decision boundaries pertain to model behavior over the whole input space.

3.3.2 Explanation Type. There are various *types* of explanation for explaining the explanandum. *Feature attribution* methods quantify the influence of particular features on the outcome. This gives insight into which features are considered important for a prediction from the perspective of the model. One may also explain a prediction by showing examples of *similar instances* with similar outcomes, or the closest data instance with a different (*counterfactual*) outcome. *Prototypical* examples may additionally provide insight in which types of instances are well represented in the data. One may also explain model behavior by providing *rules* that approximate and provide insight into the model's decision boundary.

3.3.3 Manner. Explanations may be provided in all sensory *modalities*. Currently, explanations are provided either in the *visual* (which includes explanations via text) or *auditory* (e.g. in AI assistants) modality, but other modalities might also play a role.

Whereas many of the current XAI methods provide explanations *statically*, there is increasing attention for the fact that human explanations are socially *interactive* in nature [15, 17]. For example, when a chat bot provides an explanation that is partially unclear, it is desirable that the explanation is adjusted or refined if the end user asks for clarification.

3.3.4 Strategy Type. If a model of choice has interpretable internals, one may take the *transparent by design* strategy. One can for example explain which features are important for a decision by investigating the coefficients in linear regression. In high-risk scenarios there is a strong argument for enforcing the transparent by design explanation strategy [21]. In other cases explainability has to be provided post-hoc. Some methods do this without considering model internals at all (*model-agnostic*), whereas other methods try to make model internals more interpretable (*model-specific*), for example by visualizing intermediate layers in convolutional neural networks.

4 TAGGING XAI SOLUTIONS USING ASCENT

This section describes how the proposed ontology can be used to describe XAI solutions by "tagging" them with ontology elements and illustrates this process with an example.

4.1 Method

We provide a metadata standard serialised as an OWL (Web Ontology Language) ontology[1] based on the proposed model (section 3), such that the tagging procedure is formalised and can be shared. Tagging XAI solutions according to a common ontology is a deliberate effort that promotes the FAIR application of XAI solutions [30]. Providing a serialisation of tagged XAI solutions makes these solutions easier to be unambiguously interpreted by the broader

[1] https://github.com/ajayaadhikari/ascent-xai-framework

community. Modelling the standard as an OWL ontology also facilitates future extensions to and reuse of the model. OWL stems from the semantic web and linked data paradigms, so each concept of the ASCENT framework will be identifiable through a globally unique URI that external ontologies can relate to.

Tagging according to the Use Case module additionally promotes transparency with respect to appropriate usage contexts, thereby also addressing community calls for data sheets [11] and model cards [16]. Another desirable side-effect is that manually tagging a given solution according to the Use Case module often requires an interdisciplinary effort, for example with social scientists, such that a broader societal perspective is included in the solution space from the outset.

To get rich descriptions, we require that XAI solutions are tagged with the most specific properties from the three ontology modules.[2] Each ontology element has received a dedicated property "hasX" in the serialisation, following OWL naming conventions. XAI solutions are of the type `ExplanationAlgorithm` and will require values for the "hasX" properties from the Explanation Algorithm module. The next step is to indicate relations with elements from the Use Case and AI System module. We currently define relations directly with leaf elements of the other modules for the sake of simplicity, but it is also possible to make separate instances for complex Use Case and AI System configurations and connect to those, if desired.

Indicating relations with AI System elements is relatively straightforward because they are typically static, but tagging with the Use Case module is often not. For example, even if counterfactual explanations are currently mainly connected with the explanation goal of "contesting" [29, p. 40-41], it could be connected with "justify" or "improve" in future work. That is, the absence of a tag does not imply evidence that there is no connection. Yet, it is ambiguous what absence of a tag *does* imply, because it may mean that a relation is researched but shown not to be present or even negative, or that a relation has not been researched yet. Moreover, establishing relations with Use Case elements may require validation with user research. To enrich descriptions with information about user evaluations, we have implemented additional sub-properties in the ontology to disambiguate four possible types of relations: there is research supporting 1) a *positive* relation or 2) a *negative* relation[3]; 3) or a relation has been investigated with *inconclusive* results or 4) not researched at all (*unknown*). Research gaps and possibilities for further research can be identified by filtering on inconclusive or unresearched relations. XAI solutions should be exhaustively tagged using these relation properties if detailed filtering operations are desired in a downstream application.

4.2 Example

To illustrate the tagging procedure from the perspective of a group of experts adding an XAI solution to the knowledge base, we annotated LIME [19] with elements from the three modules (figure 2).

The annotation with the AI system and Explanation Algorithm module is relatively straightforward, because it is clear what type

```
:LIMEExplanationAlgorithm
  rdf:type :ExplanationAlgorithm ;
  rdfs:label "LIME Explanation Algorithm" ;
  :hasExplanandumType :LocalDecisionBoundary ;
  :hasExplanationType :
      ↪ FeatureAttributionExplanation ;
  :hasManner :Static , :Visual ;
  :hasStrategyType :ModelAgnosticStrategy ;
  :negativeAssociationWith :Audio , :CalibrateTrust
      ↪ , :Clustering , :DimensionalityReduction ,
      ↪  :Graph , :HighRisk , :
      ↪ ReinforcementLearning , :Unsupervised , :
      ↪ Video ;
  :positiveAssociationWith :AIExpert , :
      ↪ Classification , :Consultant , :
      ↪ HighTimePressure , :Image , :
      ↪ LowTimePressure , :LowRisk , :Regression ,
      ↪ :Researcher , :Supervised , :Tabular , :
      ↪ Subject , :Text ;
  :unknownAssociationWith :Audit , :Contest , :
      ↪ DecisionSupport , :Discover , :Improve , :
      ↪ Justify , :Persuade ;
.
```

Figure 2: LIME algorithm serialised in Turtle (.ttl) according to the ontology

of models and explanations LIME supports. The properties of the Explanation Algorithm are indicated with `hasExplanandumType`, `hasExplanationType`, `hasManner`, and `hasStrategyType`. The relations with elements from the AI system and Use Case module are indicated e.g. with `positiveAssociationWith`. Notice that in this simple case, we specify relations directly with data elements of the AI System and Use Case modules. In more complex scenarios, the ontology supports creating separate Use Case instances that an XAI solution can be associated with. This may be necessary if we want to associate an XAI solution with a particular *combination* of Use Case elements. The serialisation of such a Use Case instance would similarly use the properties `hasExplanationGoal`, `hasTaskContext`, `involvesUserBackground`, and so on. These properties are all implemented in the ontology and can be found in the referenced Turtle serialisation.

The annotation with the Use Case module relies more on insights from social science research. For example, one user study showed that even though users perceive feature attribution explanations as making the prediction of the model more transparent than no explanation, they actually perform poorer when asked to predict the model's behavior after seeing the feature attribution explanation [1]. This suggests that feature attribution explanations have a negative relation with the goal of trust calibration. Research also indicates that model-agnostic strategies do not always generate faithful explanations [28]. This means LIME has a negative relation with high risk use cases in which the user needs to fully trust that the explanation reflects the true underlying reasoning of the model. Research gaps can be indicated via the `unknownAssociationWith` or `inconclusiveAssociationWith` element.

[2]Membership of supernodes is automatically inferred and does not require explicit tagging.
[3]We define a negative relation as evidence that a module node is not applicable to the XAI method.

5 CONCLUSIONS

Existing XAI survey papers generally focus on categorization of XAI solutions according to different AI System and Explanation Algorithm properties. From the application perspective in the industry, however, it is important to view XAI solutions from the space of use case needs as well. This work integrates these three aspects in the ASCENT framework for annotating XAI solutions with three modules: AI System, Use Case, and Explanation Algorithm.

The aim of the ASCENT framework is to support the FAIR usage of XAI solutions by providing rich metadata to XAI solutions, online access to the underlying ontology, modelling the ontology in a commonly used standard, and by being extendable. This is an important step towards creating a communal repository of XAI solutions that are described with metadata relevant to industry applications. By requiring XAI solutions to be connected with use case elements, the framework encourages multidisciplinary collaboration, e.g. with social scientists.

As future work, we are planning to develop a shared tool with an interface for adding, managing, and filtering ASCENT descriptions, including evaluation results of XAI solutions. Given this knowledge base, the tool could provide validated recommendations of suitable XAI solutions, given specific use case needs. It is possible that for certain use case elements, no AI solution is available or properly evaluated yet. This provides a feedback loop to the XAI research community to focus more on those research gaps.

ACKNOWLEDGMENTS

This work was part of the FATE 2021 project which is funded by the Appl.AI program within TNO.

REFERENCES

[1] Ajaya Adhikari, David M. J. Tax, et al. 2019. LEAFAGE: Example-based and Feature importance-based Explanations for Black-box ML models. In *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 1–7.

[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[3] Vaishak Belle and Ioannis Papantonis. 2020. Principles and Practice of Explainable Machine Learning. *CoRR* abs/2009.11698 (2020). arXiv:2009.11698 https://arxiv.org/abs/2009.11698

[4] Christopher M. Bishop. 2006. Pattern recognition. *Machine learning* 128, 9 (2006).

[5] Tolga Bolukbasi, Kai-Wei Chang, et al. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in neural information processing systems* 29 (2016), 4349–4357.

[6] Shruthi Chari, Oshani Seneviratne, et al. 2020. Explanation Ontology: A Model of Explanations for User-Centered AI. *CoRR* abs/2010.01479 (2020). arXiv:2010.01479 https://arxiv.org/abs/2010.01479

[7] Sanjoy Dasgupta, Nave Frost, et al. 2020. Explainable k-means clustering: theory and practice. In *XXAI Workshop. ICML*.

[8] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[9] Upol Ehsan and Mark O. Riedl. 2020. Human-centered explainable AI: Towards a reflective sociotechnical approach. In *International Conference on Human-Computer Interaction*. Springer, 449–466.

[10] European Commission. 2021. Proposal for a regulation of the European Parliament and of the Council Laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[11] Timnit Gebru, Jamie Morgenstern, et al. 2018. Datasheets for Datasets. *CoRR* abs/1803.09010 (2018). arXiv:1803.09010 http://arxiv.org/abs/1803.09010

[12] Robert R. Hoffman, Shane T. Mueller, et al. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[13] Weina Jin, Sheelagh Carpendale, et al. 2019. Bridging AI Developers and End Users: an End-User-Centred Explainable AI Taxonomy and Visual Vocabularies.

[14] Weina Jin, Jianyu Fan, et al. 2021. EUCA: A Practical Prototyping Framework towards End-User-Centered Explainable Artificial Intelligence. *CoRR* abs/2102.02437 (2021). arXiv:2102.02437 https://arxiv.org/abs/2102.02437

[15] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[16] Margaret Mitchell, Simone Wu, et al. 2018. Model Cards for Model Reporting. *CoRR* (2018). arXiv:1810.03993 http://arxiv.org/abs/1810.03993

[17] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining Explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 279–288. https://doi.org/10.1145/3287560.3287574

[18] Patrick R. Nicolas. 2015. *Scala for machine learning*. Packt Publishing Ltd.

[19] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv:1602.04938 [cs.LG]

[20] Denis Rothman. 2020. *Hands-On Explainable AI (XAI) with Python: Interpret, visualize, explain, and integrate reliable AI for fair, secure, and trustworthy AI apps*. Packt Publishing Ltd.

[21] Cynthia Rudin. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. arXiv:1811.10154 [stat.ML]

[22] Tjeerd Schoonderwoerd, Wiard Jorritsma, Mark A. Neerincx, and Karel van den Bosch. 2021. Human-Centered XAI: Developing Design Patterns for Explanations of Clinical Decision Support Systems. *International Journal of Human-Computer Studies* (2021), 102684.

[23] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.

[24] Ilaria Tiddi, Mathieu D'Aquin, and Enrico Motta. 2015. An ontology design pattern to define explanations. In *Proceedings of the 8th International Conference on Knowledge Capture, K-CAP 2015 (Proceedings of the 8th International Conference on Knowledge Capture, K-CAP 2015)*. Association for Computing Machinery, Inc. https://doi.org/10.1145/2815833.2815844 8th International Conference on Knowledge Capture, K-CAP 2015 ; Conference date: 07-10-2015 Through 10-10-2015.

[25] Nava Tintarev and Judith Masthoff. 2011. Designing and evaluating explanations for recommender systems. In *Recommender systems handbook*. Springer, 479–510.

[26] Jasper van der Waa, Elisabeth Nieuwburg, et al. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence* 291 (2021), 103404.

[27] Jasper van der Waa, Tjeerd Schoonderwoerd, et al. 2020. Interpretable confidence measures for decision support systems. *International Journal of Human-Computer Studies* 144 (2020), 102493.

[28] Mythreyi Velmurugan and Chun others Ouyang. 2021. Evaluating Fidelity of Explainable Methods for Predictive Process Analytics. In *International Conference on Advanced Information Systems Engineering*. Springer, 64–72.

[29] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

[30] Mark D. Wilkinson, Michel Dumontier, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 1 (2016), 160018. https://doi.org/10.1038/sdata.2016.18

[31] Kelvin Xu, Jimmy Ba, et al. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. PMLR, 2048–2057.

[32] Rex Ying, Dylan Bourgeois, et al. 2019. Gnn explainer: A tool for post-hoc explanation of graph neural networks. *arXiv preprint arXiv:1903.03894* (2019).