# Semantic Web Technologies for Explainable Machine Learning Models: A Literature Review

Arne Seeliger[1,2] (✉), Matthias Pfaff[1], and Helmut Krcmar[2]

[1] fortiss, Research Institute of the Free State of Bavaria associated with Technical University of Munich, Guerickestr. 25, 80805 Munich, Germany
seeliger@fortiss.org
[2] Technical University of Munich, Boltzmannstr. 3, 85748 Garching, Germany

**Abstract.** Due to their tremendous potential in predictive tasks, Machine Learning techniques such as Artificial Neural Networks have received great attention from both research and practice. However, often these models do not provide explainable outcomes which is a crucial requirement in many high stakes domains such as health care or transport. Regarding explainability, Semantic Web Technologies offer semantically interpretable tools which allow reasoning on knowledge bases. Hence, the question arises how Semantic Web Technologies and related concepts can facilitate explanations in Machine Learning systems. To address this topic, we present current approaches of combining Machine Learning with Semantic Web Technologies in the context of model explainability based on a systematic literature review. In doing so, we also highlight domains and applications driving the research field and discuss the ways in which explanations are given to the user. Drawing upon these insights, we suggest directions for further research on combining Semantic Web Technologies with Machine Learning.

**Keywords:** Semantic Web Technologies · Machine Learning · Explainability · XAI.

## 1 Introduction

Artificial Intelligence (AI) and Machine Learning (ML) techniques in particular have had tremendous success in various tasks including medical diagnosis, credit card fraud detection, or face recognition [11]. These systems, however, are often opaque and usually do not provide human-understandable explanations for their predictions [23]. This situation is problematic because it can adversely affect the understanding, trust, and management of ML algorithms [23]. While not every (benign) algorithmic decision needs to be explained in detail, explainability is necessary when dealing with incomplete problem statements including aspects of safety, ethics, or trade-offs [18]. Additionally, legal considerations of AI accountability add to the relevance of explainable decision systems [19].

The umbrella term *Explainable Artificial Intelligence* (XAI) is often used in academia to refer to a variety of approaches attempting to make ML methods explainable, transparent, interpretable, or comprehensible. Due to its relevance a plethora of research on XAI exists, including literature reviews of popular methods and techniques (see [2] or [22] for example). However, many of those approaches rely on a purely technical analysis of the black-box ML models. For such approaches Cherkassky and Dhar [14] argue that model explainability cannot be achieved. The authors further stipulate that explainability is highly dependent on the usage of domain knowledge and not data analysis alone. This idea has been adapted more recently by different authors arguing that the incorporation of Semantic Web Technologies might be a key to achieve truly explainable AI-systems [26, 27]. Since existing surveys on XAI have not explored this promising avenue of research in detail, we provide a literature-based overview of the usage of Semantic Web Technologies alongside ML methods in order to facilitate explainability. Specifically, we focus on addressing three research questions:

1. What combinations of Semantic Web Technologies and ML have been proposed to enhance model explainability?
2. Which domains of applications and tasks are especially important to this research field?
3. How are model explanations evaluated and presented to the user?

The remainder of this paper is organized as follows. Section 2 provides relevant background information pertaining to explainability of ML systems. Subsequently, Section 3 briefly describes the research design before presenting the main findings of this research. Based on these insights, implications for future research are presented in Section 4. Finally, Section 5 concludes this research.

## 2 Background and Scope of the Literature Review

Explainability of Artificial Intelligence is not a new stream of inquiry. Mueller et al. [39] analyzed the temporal development of XAI and showed that the topic has been intensively studied from the 1970s to the early 1990s within the context of Expert and Tutoring Systems. In the following two decades, only little research has been produced in the field. Recently, however, there has been a resurgence of the topic due to the interest in Machine Learning and Deep Learning [39].

Despite recent frequent publications on the topic of XAI there is no agreement upon a definition of explainability [34]. For the purpose of this survey, we follow Adadi and Berrada [2] in differentiating *interpretable* systems which allow users to study the (mathematical) mapping from inputs to outputs from *explainable* systems which provide understanding of the system's work logic. In this context, Doran et al. [17] postulate that truly explainable systems need to incorporate elements of reasoning which make use of knowledge bases in order to create human-understandablable, yet unbiased explanations. Furthermore, it is worth mentioning that interpretability or explainability not only depends on a specific model but also the knowledge and skills of its users [24].

Within the domain of ML a number of surveys address the topic of explainability and interpretability. For example, Biran and Cotton [10] review algorithmic and mathematical methods of interpretable ML models, Abdul et al. [1] focus on explanations from a human-centered perspective, and Adadi and Berrada [2] provide a holistic survey which also covers aspects of evaluation and perception. However, these studies often do not touch upon how tools such as Semantic Web Technologies might foster ML system explainability. In contrast, within the related field of Data Mining and Knowledge Discovery the interpretation of data patterns via Semantic Web and Linked Open Data has been described in a detailed survey by Ristoski and Paulheim [45]. While Data Mining, Knowledge Discovery, and ML certainly overlap in some areas, a clear overview of the combination of Semantic Technologies and Machine Learning is still missing. In this context it is worth mentioning that the scope of this review is on classical ML techniques as opposed to fields such as Inductive Logic Programming (ILP) [40]. ILP combines ideas from ML (learning from positive and negative examples) with logical programming in order to derive a set of interpretable logical rules. The interested reader can find a summary of how ontologies can be used in the ILP framework in [35]. While some researchers see ILP as a subcategory of ML (e.g. [50]), we follow Kazmi et al. [30] in differentiating the two fields and focus on more classical ML while touching upon ILP only briefly.

## 3 Explainable Machine Learning Models through Semantic Web Technologies

In this section we briefly lay out the research design of this survey before summarizing the insights of the conducted analysis. To answer the posed research questions we carried out an extensive literature review [58] by searching major academic databases including ACM Digital Library, SCOPUS, and peer-reviewed pre-prints on arXiv. The latter has been incorporated because XAI is a dynamically evolving field with a number of contributions stemming from ongoing work. We conducted a search based on keywords relating to three categories: Machine Learning, Semantic Web Technologies, and explainability.[1] The resulting list of papers was evaluated for relevance based on their abstracts and the remaining papers based on their full content. A forward and backward search [59] has been conducted to complement the list of relevant research articles.

To shed light on the first research question, we categorized the relevant models based on their usage of ML and Semantic Web Technologies. Specifically, we distinguished ML approaches along their learning rules (supervised, unsupervised, reinforcement learning) [48] and characterized the used Semantic Web Technologies by their semantic expressiveness. In doing so, we focused on the actually exploited knowledge rather than the underlying representation. For example, if a system incorporates an ontology but exclusively makes use of taxo-

[1] Search strings included but were not limited to: "machine learning" OR "deep learning" OR "data mining"; "explanation*" OR "interpret*" OR "transparen*"; "Semantic Web" OR "ontolog*" OR "background knowledge" OR "knowledge graph*"

nomical knowledge, it is categorized as a taxonomy. We followed Sarker et al. [47] in differentiating knowledge graphs from ontologies insofar that the former are usually a set of triples most often expressed using the Resource Description Framework (RDF) while the latter additionally posses type logics and are regularly expressed using Web Ontology Language (OWL). We addressed the second research question by observing the application domains and tasks of the analyzed systems. We provide answers to the third research question by describing in what form explanations are given to the user and how their quality is assessed.

### 3.1 Combining Semantic Web Technologies with Machine Learning

The results of categorizing the relevant literature along the dimensions laid out before are presented in Table 1. From a general point of view, one can observe that Semantic Web Technologies are used primarily to make two types of ML models explainable: supervised classification tasks using Neural Networks and unsupervised embedding tasks. The Semantic Web Technologies utilized alongside Neural Networks are quite diverse, while embedding methods usually incorporate knowledge graphs. Further, systems which attempt to enhance the explainability of ML systems agnostic of the underlying algorithms mainly harness ontologies and knowledge graphs. Table 1 also illustrates that only one of the reviewed articles covers reinforcement learning. In the following paragraphs we present more in-depth findings for each type of ML approach.

Concerning **supervised learning** (classification) techniques, Table 1 illustrates that Neural Networks are the dominant prediction model. The architectures proposed are manifold and include, among others, recurrent (e.g. [16, 57]) and convolutional (e.g. [13]) networks as well as autoencoders (e.g. [5, 6]). In combining these models with Semantic Web Technologies one approach is to map network inputs or neurons to classes of an ontology or entities of a knowledge graph. For example, Sarker et al. [47] map scene objects within images to classes of the Suggested Upper Merged Ontology. Based on the image classification outputted by the Neural Network, the authors run DL-Learner on the ontology to create class expressions that act as explanations. Similarly, in the work of [56], image contents are extracted as RDF triples and then matched to DBpedia via the predicate *same-concept*. In order to answer questions provided by the user about an image, the system translates each question into a SPARQL query which is run over the combined knowledge base. The results of this operation are then used to give an answer and substantiate it with further evidence that acts as an explanation. A related approach is used in [21] to explain image recognition on classes that have not been part of any training data (zero-shot learning). Furthermore, Selvaraju et al. [49] learn a mapping between individual neurons and domain knowledge. This enables the linking of a neuron's weight (importance) to semantically grounded domain knowledge. Another common approach within the supervised classification group is to utilize the taxonomical information of a knowledge base. These hierarchical relationships aid the explanation generation in different ways. For instance, Choi et al. [15] and Ma et al. [36] design attention mechanisms while authors such as Che et al. [12] and

**Table 1.** Overview of Reviewed Articles

| Author | Machine Learning Technique | | | | | | Semantic Expressiveness | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Supervised * | | Unsupervised | | Reinforcement | Multiple | | | | |
| | Neural Network | Other | Clustering | Embedding | MDP ** | | Ontology | Knowledge Graph | Taxonomy | Glossary/Lexicon |
| Aditya et al. [3] | | | | x | | | | x | | |
| Ai et al. [4] | | | | x | | | | x | | |
| Alirezaie et al. [5, 6] | x | | | | | | x | | | |
| Batet et al. [7, 8] | | | x | | | | | | x | |
| Bellini et al. [9] | | | | x | | | | x | | |
| Che et al. [12] | x | | | | | | | | x | |
| Chen et al. [13] | x | | | | | | x | | | |
| Choi et al. [15] | x | | | x | | | | | x | |
| Clos et al. [16] | x | | | | | | | | | x |
| Geng et al. [21] | x | | | | | | x | | | |
| Gusmão et al. [24] | | | | x | | | | x | | |
| Huang et al. [28] | | | | x | | | | x | | |
| Jiang et al. [29] | | x | | | | | | | x | |
| Khan et al. [31] | | | | | x | | x | | | |
| Krishnan et al. [32] | | | | | | x | x | | | |
| Liao et al. [33] | | | | x | | | | | x | |
| Ma et al. [36] | x | | | x | | | | | x | |
| Ma et al. [37] | | | | x | | | | x | | |
| McGuinness et al. [38] | | | | | | x | x | | | |
| Musto et al. [41] | | | | | | x | | x | | |
| New et al. [42] | | | | | | x | x | | | |
| Publio et al. [43] | | | | | | x | x | | | |
| Racoceanu & Capron [44] | | | | | | x | x | | | |
| Sarker et al. [47] | x | | | | | | x | | | |
| Selvaraju et al. [49] | x | | | | | | | | | x |
| Tiddi et al. [50, 51] | | | x | | | | | x | | |
| van Engelen et al. [52] | | | | | | x | | x | | |
| Wan et al. [54] | | x | | | | | | | x | |
| Wang et al. [55] | | | | x | | | | x | | |
| Wang et al. [56] | x | | | | | | x | | | |
| Wang et al. [57] | x | | | | | | | x | | |
| Yan et al. [60] | x | | | | | | x | | | |
| Zhang et al. [61] | | | | x | | | | x | | |

* Supervised learning comprises of classification approaches only because in this review regression models were only used in systems developed for multiple techniques.
** Markovian Decision Process (MDP)

Jiang et al. [29] employ model regularization based on this domain knowledge. It should be noted, however, that these systems focus more on interpretability than explainability. Since these approaches are often found in the health care domain they are more thoroughly discussed in Section 3.2.

Regarding **unsupervised learning**, we identified two groups within the reviewed literature. As shown in Table 1, a significant body of research aims at creating explainable embeddings of or with knowledge graphs. For the most part these approaches are part of some recommendation engine and are thus explained in more detail in Section 3.2. Apart from these, a smaller number of scholars strive to increase the level of interpretability or explainability for clustering algorithms. Batet et al. [7] use the taxonomical knowledge encoded in WordNet to derive a semantic similarity function which leads to more interpretable clusters. The authors present an extension to their work [8] which allows the incorporation and merging of multiple ontologies within their framework. However, no specific explanations are provided by the system as to how cluster membership of data points can be justified. Tiddi et al. [50, 51] go beyond semantic similarity functions and propose to explain clusters or data patterns (agnostic of the clustering algorithm) by traversing a knowledge graph to find commonalities among the clusters. The system, called Dedalo, uses ILP to generate candidate explanations based on the background knowledge and the given clusters. The former is built by dynamically following the URI links of the items in the data set. However, such a technique raises the question of explanation fidelity, thus asking whether the given explanation actually agrees with the underlying predictive model.

As stated above, only one reviewed system aims at explaining **reinforcement learning**. In this research [31] the authors utilize an ontology to incorporate domain knowledge into the explanation process of an MDP recommendation system. The ontology is used to provide information which is not available from the data alone and to perform inference to create rules which limit the number of actions recommended. Finally, Semantic Web Technologies such as ontologies can be used to aid explainability and interpretability from a more general and **model agnostic** point of view. Along these lines, Krishnan et al. [32] design an explainable personal assistant that uses an ontology to dynamically grow a knowledge base, interact with other modules, and perform reasoning. In addition, Racoceanu and Capron [44] design a medical imaging platform which provides decision reproducibility and traceability powered by an ontology. Even more general, some authors propose ontologies or interlingua to declaratively represent aspects and dimensions of explainability. For instance, McGuinness et al. [38] create three ontologies with concepts and relation about data provenance, trust, and justifications, thus offering an explanation infrastructure. Similarly, by constructing an ML schema, Publio et al. [43] aim at exposing the semantics of such systems which can positively affect model explainability.

Lastly, we want to highlight another insight relating to the performance of the explainable systems. It is worth noting that in using Semantic Web Technologies alongside ML algorithms, explainability is not raised at the cost of performance. Rather, the reviewed systems often achieve state-of-the-art performance in their respective tasks. This is particularly notable because these results exemplify how to overcome the often assumed trade-off between ML accuracy and interpretability by the means of structure and logic [46].

### 3.2 Domains and Applications

The combinations of ML algorithms and Semantic Web Technologies are also driven by the respective application domains and tasks to be accomplished. Table 2 provides an overview of the most frequent domains and tasks of the reviewed systems. Regarding the former, it becomes apparent that – while many systems are developed agnostic of a specific domain – health care is a strong driver for interpretable ML systems. Regarding the tasks of the reviewed systems, we found the recommendation task and image analysis to be of great importance. For brevity we limit the following paragraphs to the health care domain and the recommendation task.

**Table 2.** Selected Domains of Application and Tasks

| | Tasks and Domains | Authors |
|---|---|---|
| Domains | General | [3], [16], [24], [28], [38], [43], [47], [50–52], [55, 56], [61] |
| | Health Care | [12], [15], [29], [36], [42], [44], [54], [60] |
| | Entertainment | [9], [41], [57] |
| | Commercial | [4], [33], [37] |
| Tasks | Recommendation | [4], [9], [28], [31], [37], [41], [55], [57] |
| | Image Annotation or Classification | [5, 6], [21], [44], [47], [49], [60] |
| | Transfer or Zero-Shot Learning | [13], [21], [49] |
| | Knowledge Base Completion | [24], [52], [61] |
| | Diagnosis Prediction | [12], [15], [36] |
| | Visual Question Answering | [3], [56] |

Note: Multiple selections possible.

Systems in the domain of **health care** often combine classification tasks such as diagnosis prediction with taxonomical knowledge found in medical diagnosis codes or medical ontologies. For instance, Jiang et al. [29] use the hierarchical information of the International Classification of Diseases (ICD) to introduce a regularization penalty to their logistic regression which produces a sparse model where non-zero features tend to be localized within a limited number of subtrees instead of being scattered across the entire hierarchy. This kind of feature weighting might make the algorithmic prediction process more explicit (interpretability), but it does not provide explanations and justification for laymen (e.g. patients). Similarly, Chen et al. [12] incorporate hierarchical ICD knowledge in a Neural Network architecture to regularize the output layer of the network and learn clinically relevant features. Yan et al. [60] use hierarchical relationships within an ontology to expand a set of medical labels by inferring missing parent labels. For example, the label "right mid lung" is expanded to "right lung", "lung", and "chest". The authors also utilize exclusive relationships between labels to learn hard cases and improve accuracy. When making predictions on medical images, their system is able to provide input examples similar to the given model output as prediction evidence. Finally, KAME [36] is a diagnosis prediction system inspired by [15] which uses medical ontologies to learn (embedded) representations of medical codes and their parent codes. These are

then utilized to learn input representations of patient data which are fed into a Neural Network architecture. The authors exploit an attention mechanism which learns weights that allow to interpret the importance of different pieces of knowledge. Summing up, within the domain of health care many interpretable ML models have been proposed. These mainly use taxonomical knowledge to aid performance and interpretability. The reason for the relative abundance of such systems in the health care domain stems from the high stakes characteristics of the field as well as the existence of different medical ontologies.

Due to their extensive use of knowledge graphs, **recommendation systems** are an important branch of research in the reviewed field. More specifically, these systems commonly combine embedding models with knowledge graphs. For example, Bellini et al. [9] inject the DBpedia knowledge graph into an autoencoder network which is constructed to mirror the structure of the knowledge base. After training such a system for each user, the learned weights map to explicit semantic concepts from the knowledge graph and user-specific explanations can be generated based on these insights. Another special case of embedding is RippleNet [55] where the triples of a constructed knowledge graph (based on Microsoft Satori) are iteratively compared to the embeddings and then propagated. This way the path from a user's history to a recommended item can be used as an explanation for the recommendation. Further, there are approaches which use Semantic Web Technologies agnostic of the underlying recommendation algorithm. One such system is ExpLOD [41] which makes use of the Linked Open Data paradigm. The framework first maps liked items and recommended items into a knowledge base such as DBpedia, then builds a graph, ranks the properties in this graph based on relevance, and finally creates a natural language explanation from the top properties retrieved. While being model agnostic, the issue of explanation fidelity can be raised again here because the given explanation might not correspond to the actual underlying model process. Finally, it is worth mentioning that explainability in recommender systems is mainly driven from a user-centric perspective with the aim to increase user satisfaction and acceptance.

### 3.3 Explanation Forms and Evaluation

The conducted analysis revealed that the presentation and form of the given explanations is highly diverse – even within similar domains or prediction tasks. For example, some scholars combine different types of explanations (e.g. visual and textual [49]) in order to increase explainability while others provide only minimal explanation towards the user (e.g. [7] or [52]). Moreover, only few authors present explanations in natural language. For instance, Musto et al. [41] incorporate a dedicated natural language generator into their recommendation algorithm. The authors utilize a template-based approach which is also used by other authors [4, 31]. A more frequently employed explanation form consists of textual (semi)-logical or rule-like notation. Further, explanations are usually designed to optimally justify correct model output. One deviation from this is the work of Alirezaie et al. [5, 6] where the *errors* of a Neural Network image classifier are explained by performing ontological reasoning upon objects of a

scene. To illustrate the range of explanation forms used, Table 3 provides selected examples of textual explanations encountered in this review. Apart from the ambiguity of the term explainability, one potential reason for this diversity includes the relevancy of an explanation for a given system: While in most reviewed cases, explainability is an explicit goal, in a subset of models, explainability is treated as a secondary goal and Semantic Web Technologies are used to primarily address other issues such as data sparseness (e.g. [15]).

**Table 3.** Examples of Textual Explanations

| Author | Task | Example Explanation |
|---|---|---|
| Bellini et al. [9] | Recommendation of a movie | **Prediction:**<br>Terminator 2<br>**Explanation:**<br>We guess you would like to watch Terminator 2: Judgment Day (1991) more than Transformers: Revenge of the Fallen (2009) because you may prefer:<br>• (subject) 1990s science fiction films [...]<br>over:<br>• (subject) Films set in Egypt [...] |
| Gusmão et al. [24] | Knowledge graph completion (triple prediction) | **Prediction:**<br>Head: francis_ii_of_the_two_sicilies,<br>Relation: RELIGION,<br>Tail: roman_catholic_church<br>**Explanation:**<br>#1: parents, religion<br>#2: spouse$^{-1}$, religion [...] |
| Selvaraju et al. [49] | Image classification of an animal | **Prediction:**<br>Yellow-headed blackbird<br>**Explanation:**<br>has_eye_color = black, has_underparts_color = white, has_belly_color = white, has_breast_color = white, has_breast_pattern = solid |
| Zhang et al. [61] | Knowledge graph completion (link prediction) | **Prediction:**<br>World War I – entity envolved – German Empire<br>**Explanation:**<br>World War I – commanders – Erich Ludendorff<br>Erich Ludendorff – commands – German Empire<br>Supported by:<br>Falkland Wars – entities envolved – United Kingdom<br>Falkland Wars – commanders – Margaret Thatcher<br>Margaret Thatcher – commands – United Kingdom |

Note: Some explanations have been shortened for legibility as indicated by square brackets.

Furthermore, we found most systems to offer rather static explanations without much user interaction. In this context, the work of Liao et al. [33] is an exception as the proposed recommendation system enables user-feedback on human-interpretable domain concepts. Moreover, looking into the future, Sarker et al. [47] envision their explanation tool for image classification to be used in an interactive human-in-the-loop system where a human monitor can correct algorithmic decisions based on the given explanations. On the whole, however, we notice a lack of user-adaptive or interactive explanation approaches in the reviewed literature.

Finally, when it comes to evaluating the goodness of the explanations, only few authors go beyond a subjective assessment of the proposed system. Bellini et al. [9], for instance, perform an evaluation of their knowledge-aware autoencoder recommendation system by conducting A/B testing with 892 volunteers. Similarly, Musto et al. [41] designed a user study in which 308 subjects filled out a questionnaire involving questions such as *"I understood why this movie was recommended to me"*. Through this evaluation, the authors gain further insights into different aspects of how their explanation system affects end users. Other authors propose more quantitative evaluation metrics to determine the goodness of the given explanations. Zhang et al. [61] explain their link predictions by finding patterns within a knowledge graph which are similar to the predicted ones (see Table 3) and measure explanation reliability by the number of similar patterns found. Further, Jiang et al. [29] measure the interpretability of their predictive system by quantifying the sparseness of their linear model while taking into account the taxonomical structure of their data. Overall, from these findings it becomes obvious that there is no accepted standard for evaluating explanations within XAI.

## 4 Trends for Future Research

Based on our review of the relevant literature we articulate opportunities and challenges for future research in the field. We generate these insights based on our analysis and comparison among all reviewed papers as well as on the basis of the challenges put forward within each of the articles.

### 4.1 Semantic Web Technologies for Explainability

The combination of Semantic Web Technologies and ML offers great potential for facilitating explainable models. We identified the matching of ML data with knowledge base entities – which has been called *knowledge matching* [21] – as one central challenge which needs to be overcome by future research. Specifically, automated and reliable methods for knowledge matching are required. In this context, Wang et al. [56] suggest string matching between identified objects and ontology classes and Liao et al. [33] propose to mine concepts and relationships automatically from online sources. Further research in this area as well as related fields like semantic annotation are needed to enable effective and efficient knowledge matching.

Moreover, we found a certain concentration on specific ML techniques and Semantic Web Technologies. More work needs to be conducted on explainable reinforcement learning and clustering. In this context, we also note that the work across different disciplines and tasks still remains somewhat isolated even though concepts like linked data provide the tools for integrating various domains. Some existing research acknowledges the need to extend the range of tasks performed by explainable systems [12] or their domains of application [32]. Other authors envision the use of more data [60] or more complex background knowledge [41,

42, 47]. Hence, the areas of ontology or knowledge graph learning as well as knowledge base matching play an important role in accomplishing this goal. Future work will therefore need to find ways to mitigate the potential lack of data interconnectedness and the increased complexity of such systems.

Finally, we highlight the need for future work to aim for truly explainable systems which incorporate reasoning and external knowledge that is human-understandable. To achieve this goal, future explanation systems need to ensure that the explanations given are truthful to the underlying ML algorithm. Further, such approaches should be able to explain not only how an output relates to some representation of interest but also how this representation has been obtained. For example, it is not enough to justify that a human face has been detected by stating that eyes, mouth, and nose were recognized and that these features are part of a human face (e.g. inferred via ontology). A truly explainable system should also be able to explain why these features have been recognized. This point relates to the question of user interaction, which is discussed below.

### 4.2 Human-Centric Explanations

Since explanations are forms of social interactions [2], their efficacy and quality depend to a large extent on their intelligibility and comprehensibility as perceived by the user. In other words, an explanation is only useful if the user is able to understand it. In this review we have shown that the form and appearance of explanations differs significantly among current systems and many of those do not provide explanations in natural language. Therefore, we believe that the field of Natural Language Processing (NLP) and Natural Language Generation (NLG) in particular offers a useful starting point. For example, Vougiouklis et al. [53] generate natural texts from Semantic Web triples using Neural Networks. Moreover, Ell et al. [20] translate SPARQL queries to English text that is understandable by non-experts. More generally, the field of (Visual) Question Answering can be a source of inspiration since questions and answers are usually given in natural language [56].

Additionally, we believe that explanations need to be adaptive and interactive in order to generate the greatest benefit for the user. Structured knowledge bases could allow users to scrutinize and interact with explanations in various forms. For example, user could browse among different possible explanations or drill down on a specific explanation to extract more specific reasons that contributed to a prediction. Khan et al. [31] envision a system that allows for such follow up questions. Similarly, Bellini et al. [9] plan to incorporate the possibility for users to correct their system in a continuous loop. As described above, Sarker et al. [47] also regard this course of action as an important task for future studies. However, there seems to be no consensus regarding the actual mode of interaction. In order to find optimal ways of presenting and interacting with explanations, future research needs to incorporate findings from a greater variety of research fields. Existing studies [1, 2] show that there is a growing body of diverse and interdisciplinary work addressing the question of human-understandable explanations that can be leveraged in this context.

### 4.3 Common Grounds for Evaluation

We believe that meaningful progress in the field of XAI is not only dependent on novel explanation algorithms but also on common grounds for model evaluation and comparison. In light of this, Doshi-Velez and Kim [18] put forward the need for a shared language relating to factors of ML explainability. We have shown that Semantic Web Technologies can help in creating such a common lingua. Future work, however, needs to proof how to utilize such constructs effectively in the context of explainability. Another way forward could be to develop and rely on standard design patterns for combining ML with Semantic Web Technologies. The work of van Harmelen and ten Teije [25] already provides a collection of patterns for such hybrid systems. Moreover, common evaluation criteria need to be established so that subjective assessments of model explainability can be replaced by more rigorous practices.

## 5 Conclusion

Explainability and interpretability have become an essential requirement for many ML systems. In this work, through an extensive literature review, we have shown that the connection between ML and Semantic Web Technologies can yield exciting opportunities regarding model explainability. We discussed the most prevalent approaches within supervised and unsupervised learning and highlighted how the domain of health care and the recommendation task are important drivers of the research field. The literature analysis further revealed that prediction performance is not reduced but often increased by incorporating background knowledge within the ML paradigm. Finally, we provided examples of specific forms of explanations including natural language and rule-like statements. At the same time, we highlighted that meaningful progress in the reviewed field also relies on advances in a number of research challenges. These include technical questions like automated ways of knowledge matching or progress in knowledge base learning. Other challenges concern the development of adaptive and interactive systems. Lastly, more rigorous evaluation strategies need to be devised by future research. We believe that tackling these questions and further exploring the combination of structured knowledge, reasoning, and Machine Learning can pave the way to truly explainable systems.

## References

1. Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y., Kankanhalli, M.: Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 582:1–582:18. CHI '18, ACM, New York, NY, USA (2018)
2. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)

3. Aditya, S., Yang, Y., Baral, C.: Explicit reasoning over end-to-end neural architectures for visual question answering. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans, Louisiana, USA (2018)
4. Ai, Q., Azizi, V., Chen, X., Zhang, Y.: Learning heterogeneous knowledge base embeddings for explainable recommendation. Algorithms **11**(9), 137 (2018)
5. Alirezaie, M., Längkvist, M., Sioutis, M., Loutfi, A.: A symbolic approach for explaining errors in image classification tasks. In: IJCAI Workshop on Learning and Reasoning. Stockholm, Sweden (2018)
6. Alirezaie, M., Längkvist, M., Sioutis, M., Loutfi, A.: Semantic Referee: A neural-symbolic framework for enhancing geospatial semantic segmentation. Semantic Web Journal (2019)
7. Batet, M., Valls, A., Gibert, K.: Performance of ontology-based semantic similarities in clustering. In: Proceedings of the 10th International Conference on Artificial Intelligence and Soft Computing. pp. 281–288. Springer, Berlin, Heidelberg (2010)
8. Batet, M., Valls, A., Gibert, K., Sánchez, D.: Semantic clustering using multiple ontologies. In: Artificial Intelligence Research and Development - Proceedings of the 13th International Conference of the Catalan Association for Artificial Intelligence. pp. 207–216. IOS Press, Amsterdam, The Netherlands (2010)
9. Bellini, V., Schiavone, A., Di Noia, T., Ragone, A., Di Sciascio, E.: Knowledge-aware autoencoders for explainable recommender systems. In: Proceedings of the 3rd Workshop on Deep Learning for Recommender Systems. pp. 24–31. DLRS 2018, ACM, New York, NY, USA (2018)
10. Biran, O., Cotton, C.: Explanation and justification in machine learning: A survey. In: Proceedings of the IJCAI-17 Workshop on Explainable AI (XAI). pp. 8–13. Melbourne, Australia (2017)
11. Brynjolfsson, E., Mitchell, T.: What can machine learning do? Workforce implications. Science **358**(6370), 1530–1534 (2017)
12. Che, Z., Kale, D., Li, W., Bahadori, M.T., Liu, Y.: Deep computational phenotyping. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 507–516. KDD '15, ACM, New York, NY, USA (2015)
13. Chen, J., Lecue, F., Pan, J.Z., Horrocks, I., Chen, H.: Knowledge-based transfer learning explanation. In: Sixteenth International Conference on Principles of Knowledge Representation and Reasoning. pp. 349–358. Tempe, AZ, USA (2018)
14. Cherkassky, V., Dhar, S.: Interpretation of black-box predictive models. In: Measures of Complexity, pp. 267–286. Springer, New York (2015)
15. Choi, E., Bahadori, M.T., Song, L., Stewart, W.F., Sun, J.: GRAM: Graph-based attention model for healthcare representation learning. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 787–795. KDD '17, ACM, New York, NY, USA (2017)
16. Clos, J., Wiratunga, N., Massie, S.: Towards explainable text classification by jointly learning lexicon and modifier terms. In: IJCAI-17 Workshop on Explainable AI (XAI). pp. 19–23. Melbourne, Australia (2017)
17. Doran, D., Schulz, S., Besold, T.R.: What does explainable AI really mean? A new conceptualization of perspectives. In: Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 co-located with 16th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2017). Bari, Italy (2017)
18. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)

19. Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Schieber, S., Waldo, J., Weinberger, D., Wood, A.: Accountability of AI under the law: The role of explanation. Berkman Center Research Publication Forthcoming; Harvard Public Law Working Paper No. 18-07 (2017)
20. Ell, B., Harth, A., Simperl, E.: SPARQL query verbalization for explaining semantic search engine queries. In: Presutti, V., d'Amato, C., Gandon, F., d'Aquin, M., Staab, S., Tordai, A. (eds.) The Semantic Web: Trends and Challenges, pp. 426–441. Springer, Cham (2014)
21. Geng, Y., Chen, J., Jimenez-Ruiz, E., Chen, H.: Human-centric transfer learning explanation via knowledge graph. In: AAAI Workshop on Network Interpretability for Deep Learning. Honolulu, HI, USA (2019)
22. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of Machine Learning. In: 5th International Conference on Data Science and Advanced Analytics (DSAA). pp. 80–89. IEEE, Turin, Italy (2018)
23. Gunning, D.: Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency (DARPA) (2017)
24. Gusmão, A.C., Correia, A.H.C., De Bona, G., Cozman, F.G.: Interpreting embedding models of knowledge bases: A pedagogical approach. In: ICML Workshop on Human Interpretability in Machine Learning (WHI). Stockholm, Sweden (2018)
25. van Harmelen, F., ten Teije, A.: A boxology of design patterns for hybrid learning and reasoning systems. Journal of Web Engineering **18**(1), 97–124 (2019)
26. Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B.: What do we need to build explainable AI systems for the medical domain? arXiv preprint arXiv:1712.09923 (2017)
27. Holzinger, A., Kieseberg, P., Weippl, E., Tjoa, A.M.: Current advances, trends and challenges of machine learning and knowledge extraction: From machine learning to explainable AI. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) Machine Learning and Knowledge Extraction. pp. 1–8. Springer, Cham (2018)
28. Huang, J., Zhao, W.X., Dou, H., Wen, J.R., Chang, E.Y.: Improving sequential recommendation with knowledge-enhanced memory networks. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 505–514. SIGIR '18, ACM, New York, NY, USA (2018)
29. Jiang, J., Chandola, V., Hewner, S.: Tree-based regularization for interpretable readmission prediction. In: AAAI 2019 Spring Symposium on Combining Machine Learning with Knowledge Engineering (AAAI-MAKE). Palo Alto, CL, USA (2019)
30. Kazmi, M., Schller, P., Saygn, Y.: Improving scalability of inductive logic programming via pruning and best-effort optimisation. Expert Systems with Application **87**(C), 291–303 (2017)
31. Khan, O.Z., Poupart, P., Black, J.P.: Explaining recommendations generated by MDPs. In: Proceedings of the Third International Conference on Explanation-aware Computing. EXACT'08, vol. 391, pp. 13–24. CEUR-WS, Aachen, Germany (2008)
32. Krishnan, J., Coronado, P., Reed, T.: Seva: A systems engineer's virtual assistant. In: AAAI 2019 Spring Symposium on Combining Machine Learning with Knowledge Engineering (AAAI-MAKE). Palo Alto, CL, USA (2019)
33. Liao, L., He, X., Zhao, B., Ngo, C.W., Chua, T.S.: Interpretable multimodal retrieval for fashion products. In: Proceedings of the 26th ACM International Conference on Multimedia. pp. 1571–1579. MM '18, ACM, New York, NY, USA (2018)
34. Lipton, Z.C.: The mythos of model interpretability. Queue **16**(3), 30:31–30:57 (2018)

35. Lisi, F.A., Esposito, F.: On ontologies as prior conceptual knowledge in inductive logic programming. In: Berendt, B., Mladenič, D., de Gemmis, M., Semeraro, G., Spiliopoulou, M., Stumme, G., Svátek, V., Železný, F. (eds.) Knowledge Discovery Enhanced with Semantic and Social Information, pp. 3–17. Springer, Berlin, Heidelberg (2009)
36. Ma, F., You, Q., Xiao, H., Chitta, R., Zhou, J., Gao, J.: KAME: Knowledge-based attention model for diagnosis prediction in healthcare. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. pp. 743–752. CIKM '18, ACM, New York, NY, USA (2018)
37. Ma, W., Zhang, M., Cao, Y., Jin, W., Wang, C., Liu, Y., Ma, S., Ren, X.: Jointly learning explainable rules for recommendation with knowledge graph. In: The World Wide Web Conference. pp. 1210–1221. WWW '19, ACM, New York, NY, USA (2019)
38. McGuinness, D.L., Ding, L., Da Silva, P.P., Chang, C.: PML 2: A modular explanation interlingua. In: AAAI 2007 Workshop on Explanation-aware Computing. pp. 49–55. Vancouver, Canada (2007)
39. Mueller, S.T., Hoffman, R.R., Clancey, W.J., Emrey, A., Klein, G.: Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. arXiv preprint arXiv:1902.01876 (2019)
40. Muggleton, S., Raedt, L.D.: Inductive logic programming: Theory and methods. Journal of Logic Programming **19**(20), 629–679 (1994)
41. Musto, C., Narducci, F., Lops, P., De Gemmis, M., Semeraro, G.: ExpLOD: A framework for explaining recommendations based on the Linked Open Data Cloud. In: Proceedings of the 10th ACM Conference on Recommender Systems. pp. 151–154. RecSys '16, ACM, New York, NY, USA (2016)
42. New, A., Rashid, S.M., Erickson, J.S., McGuinness, D.L., Bennett, K.P.: Semantically-aware population health risk analyses. In: Machine Learning for Health (ML4H) Workshop at NeurIPS. Montreal, Canada (2018)
43. Publio, G.C., Esteves, D., Ławrynowicz, A., Panov, P., Soldatova, L., Soru, T., Vanschoren, J., Zafar, H.: ML Schema: Exposing the semantics of machine learning with schemas and ontologies. In: ICML 2018 Workshop on Reproducibility in Machine Learning. Stockholm, Sweden (2018)
44. Racoceanu, D., Capron, F.: Towards semantic-driven high-content image analysis: An operational instantiation for mitosis detection in digital histopathology. Computerized Medical Imaging and Graphics **42**, 2–15 (2015)
45. Ristoski, P., Paulheim, H.: Semantic web in data mining and knowledge discovery. Web Semantics: Science, Services and Agents on the World Wide Web **36**(C), 1–22 (2016)
46. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019)
47. Sarker, M.K., Xie, N., Doran, D., Raymer, M., Hitzler, P.: Explaining trained neural networks with Semantic Web Technologies: First steps. In: Proceedings of the Twelfth International Workshop on Neural-Symbolic Learning and Reasoning (NeSy). London, UK (2017)
48. Sathya, R., Abraham, A.: Comparison of supervised and unsupervised learning algorithms for pattern classification. International Journal of Advanced Research in Artificial Intelligence **2**(2) (2013)

49. Selvaraju, R., Chattopadhyay, P., Elhoseiny, M., Sharma, T., Batra, D., Parikh, D., Lee, S.: Choose your neuron: Incorporating domain knowledge through neuron-importance. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. pp. 540–556. Springer, Cham (2018)
50. Tiddi, I., d'Aquin, M., Motta, E.: Dedalo: Looking for clusters explanations in a labyrinth of linked data. In: Presutti, V., d'Amato, C., Gandon, F., d'Aquin, M., Staab, S., Tordai, A. (eds.) The Semantic Web: Trends and Challenges. pp. 333–348. Springer, Cham (2014)
51. Tiddi, I., d'Aquin, M., Motta, E.: Data patterns explained with linked data. In: Bifet, A., May, M., Zadrozny, B., Gavalda, R., Pedreschi, D., Bonchi, F., Cardoso, J., Spiliopoulou, M. (eds.) Machine Learning and Knowledge Discovery in Databases. pp. 271–275. Springer, Cham (2015)
52. Van Engelen, J.E., Boekhout, H.D., Takes, F.W.: Explainable and efficient link prediction in real-world network data. In: Boström, H., Knobbe, A., Soares, C., Papapetrou, P. (eds.) Advances in Intelligent Data Analysis XV. pp. 295–307. Springer, Cham (2016)
53. Vougiouklis, P., Elsahar, H., Kaffee, L.A., Gravier, C., Laforest, F., Hare, J., Simperl, E.: Neural Wikipedian: Generating textual summaries from knowledge base triples. Journal of Web Semantics **52-53**, 1 – 15 (2018)
54. Wan, S., Mak, M.W., Kung, S.Y.: Mem-mEN: Predicting multi-functional types of membrane proteins by interpretable elastic nets. IEEE/ACM Transactions on Computational Biology and Bioinformatics **13**(4), 706–718 (2016)
55. Wang, H., Zhang, F., Wang, J., Zhao, M., Li, W., Xie, X., Guo, M.: Ripplenet: Propagating user preferences on the knowledge graph for recommender systems. In: Proceedings of the 27th International Conference on Information and Knowledge Management. pp. 417–426. CIKM '18, ACM, New York, NY, USA (2018)
56. Wang, P., Wu, Q., Shen, C., Dick, A., Van Den Henge, A.: Explicit knowledge-based reasoning for visual question answering. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. pp. 1290–1296. IJCAI'17, AAAI Press (2017)
57. Wang, X., Wang, D., Xu, C., He, X., Cao, Y., Chua, T.S.: Explainable reasoning over knowledge graphs for recommendation. AAAI (2019)
58. Webster, J., Watson, R.T.: Analyzing the past to prepare for the future: Writing a literature review. MIS Quarterly pp. xiii–xxiii (2002)
59. Wohlin, C.: Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering. pp. 38:1–38:10. EASE '14, ACM, New York, NY, USA (2014)
60. Yan, K., Peng, Y., Sandfort, V., Bagheri, M., Lu, Z., Summers, R.M.: Holistic and comprehensive annotation of clinically significant findings on diverse ct images: Learning from radiology reports and label ontology. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA (2019)
61. Zhang, W., Paudel, B., Zhang, W., Bernstein, A., Chen, H.: Interaction embeddings for prediction and explanation in knowledge graphs. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. pp. 96–104. WSDM '19, ACM, New York, NY, USA (2019)