On the Multiple Roles of Ontologies in Explanations for Neuro-symbolic AI

Roberto Confalonieri [a,*] and Giancarlo Guizzardi [b]

[a] *Department of Mathematics 'Tullio Levi-Civita', University of Padua, Padova, Italy*
E-mail: roberto.confalonieri@unipd.it

[b] *Department of Semantics, Cybersecurity & Services (SCS), University of Twente, Enschede, The Netherlands*
E-mail: g.guizzardi@utwente.nl

Abstract. There has been a renewed interest in symbolic AI in recent years. Symbolic AI is indeed one of the key enabling technologies for the development of neuro-symbolic AI systems, as it can mitigate the limited capabilities of black box deep learning models to perform reasoning and provide support for explanations. This paper discusses the different roles that explicit knowledge, in particular ontologies, can play in drawing intelligible explanations in neuro-symbolic AI. We consider three main perspectives in which ontologies can contribute significantly, namely reference modelling, common-sense reasoning, and knowledge refinement and complexity management. We overview some of the existing approaches in the literature, and we position them according to these three proposed perspectives. The paper concludes by discussing some open challenges related to the adoption of ontologies in explanations.

Keywords: Neuro-symbolic AI, Explanations, Applied Ontologies

1. Introduction

The limited capability of deep learning systems to perform abstraction, reasoning and to support explainability have prompted a heated debate about the value of symbolic AI in contrast to neural computation [1, 2]. Researchers identified the need for the development of hybrid systems, that is, systems that integrate neural models with logic-based approaches to provide AI systems capable of transferring learning and bridge the gap between lower-level information processing (e.g., for efficient data acquisition and pattern recognition) and higher-level abstract knowledge (e.g., for general reasoning and explainability).

Neuro-symbolic AI [1, 2] seeks to offer a principled way of integrating learning and reasoning, by aiming at establishing correspondences between neural models and logical representations [1]. The development of hybrid systems is widely seen as one of the major challenges facing AI today [2]. Indeed, there is no consensus on how to achieve this, with proposed techniques in the literature ranging from inductive logic programming [3], logical tensor networks [4], Markov logic networks [5] and logical neural networks [6], to name a few. What seems widely accepted is that knowledge representation—in its many incarnations—is a key asset to enact such hybrid systems.

The benefits of adopting explicit knowledge have also been recognized in Explainable AI (XAI) [7, 8]. XAI research focuses on the development of methods and techniques that seek to provide explanations of how deep learning and machine learning models, which are deemed *black boxes*, make decisions. Explainability is crucial

*Corresponding author. E-mail: roberto.confalonieri@unipd.it.

for designers and developers to improve system robustness, enable diagnostics for bias prevention, address fairness and discrimination concerns, and foster trust among users regarding the decision-making process. Moreover, explicability is considered one of the essential ethical dimensions for autonomous systems [9].

Current approaches to XAI focus on the mechanistic aspects of explanations, that is, generating explanations that describe *how* decisions are made [10]. One typical approach is the extraction of local and global post-hoc explanations that approximate the behaviour of a black model model by means of an interpretable proxy. However, several works in this field have stressed the fact that these approaches do not provide well-founded justifications and logical reasons for *why* decisions are made, often making these explanations not understandable by users and not useful for them [11]. To this end, using explicit knowledge in explanations can support reasoning scenarios in which formal approaches to human and common-sense reasoning can be exploited to draw explanations that are closer to the way humans think [12].

Explicit knowledge is thus of paramount importance for the development of hybrid systems and the provision of intelligible explanations. This position paper explores the role of explicit knowledge representation artifacts (i.e., symbolic structures), such as ontologies and knowledge graphs in neuro-symbolic AI, particularly supporting explainability and the generation of human-understandable explanations.

The rest of the paper is organized as follows. Firstly, we provide a concise introduction to ontologies and their typical conceptualisation and formalisation in computer science. Next, we present and elaborate on three perspectives that demonstrate how formal knowledge can contribute to the development of intelligible explanations. In support of these perspectives, we summarise existing works that align with each viewpoint. Throughout the paper we provide several examples that illustrate the main concepts. Finally, we outline several future challenges associated with ontologies and explanations in neuro-symbolic AI.

2. Ontologies and their Role in Explanations

Several recent surveys and position papers [7, 8, 13, 14] emphasize the importance of designing explainability solutions that cater to diverse purposes and stakeholders, and highlight the limitations of existing approaches in supporting comprehensive human-centric Explainable AI. Current methods predominantly concentrate on specific types and formats of explanations, primarily focusing on the mechanistic aspects that explain *how* decisions are made. However, they often overlook the need to address the more fundamental question of *why* decisions are made, such as offering causal and counterfactual explanations [11].

In addition, it is important to note that the explanations are based on background knowledge. This background knowledge encompasses both the decision being explained and the intended recipient of the explanation. Explainability techniques must bridge the communication gap between the AI system and the users of the explanations, tailoring their outputs to different user groups. Consequently, if the explanations are not easily understandable by the users, they may be compelled to seek additional knowledge to obtain reliable insights and avoid drawing false conclusions.

The use of explicit knowledge, such as ontologies, knowledge graphs, or other forms of structured formal knowledge, can potentially help to bridge these gaps [15]. There has been a renewed interest in incorporating explicit knowledge into AI in recent years. Ontologies and knowledge graphs have been successfully applied in various domains, including knowledge-aware news recommender systems [16], semantic data mining and knowledge discovery [17, 18], as well as natural language understanding [19, 20]. These applications highlight the value of leveraging explicit knowledge to enhance Machine Learning systems and address the challenges of explainability and user comprehension.

In the subsequent sections, we will commence by offering a concise introduction to ontologies, as they are commonly understood in AI, in particular, in areas such as knowledge representation and the Semantic Web. Subsequently, we will explore the potential role that ontologies can assume in explainability.

TBox:

$\mathsf{Person} \sqsubseteq \top$	,	$\mathsf{Title} \sqsubseteq \top$
$\mathsf{Researcher} \sqsubseteq \mathsf{Person} \sqcap \exists \mathsf{hasTitle.PhD}$	,	$\mathsf{PhD} \sqsubseteq \mathsf{Title}$
$\mathsf{Professor} \sqsubseteq \mathsf{Researcher} \sqcap \exists \mathsf{hasTitle.Tenured}$	,	$\mathsf{Tenured} \sqsubseteq \mathsf{Title}$

ABox:

$\mathsf{Researcher}(\mathsf{Bob})$	,	$\mathsf{Professor}(\mathsf{Mike})$
$\mathsf{CollaboratesWith}(\mathsf{Bob}, \mathsf{Mike})$	,	$\mathsf{SupervisedBy}(\mathsf{Bob}, \mathsf{Mike})$

Fig. 1. An ontology excerpt for the university domain formalised in DL. $\sqsubseteq$ is the subsumption relation, $\sqcap$ is conjunction, and $\exists$ is the existential relation. $\top$ is the top concept in the ontology. The TBox axioms state that 'every researcher is a person with a PhD title', and 'every professor is a researcher with a tenured position'. The Abox axioms state that Bob is a researcher, Mike is a professor, Bob and Mike collaborate, and Bob is supervised by Mike.

2.1. *Ontologies*

Within the realm of computer science, ontologies serve as a formal method for representing the structure of a specific domain. They capture the essential entities and relationships that arise from observing and understanding the domain, enabling its comprehensive description [21].

To illustrate the concept further, let us consider a simple conceptualisation of the domain of the university and its employees. In an ontology, the entities within this domain can be organised into concepts and relations using unary and non-unary (typically, binary) predicates. At the core of the ontology there is a hierarchy of concepts, known as a taxonomy. For instance, if our focus is on academic roles, we might have relevant concepts such as *Person*, *Researcher* and *Professor*. In this case, *Person* would be a super-concept of *Researcher* and *Professor*. Additionally, a relevant relation could be *CollaboratesWith* which captures the connections between individuals. Within the ontology, a specific researcher employed by the university would be an instance of the corresponding concept, linking them to a broader domain structure.

From a formal representation perspective, an ontology can be defined as a collection of axioms formulated using a suitable logical language. These axioms serve to express the intended semantics of the ontology, providing a conceptual framework for understanding a specific domain. It is worth noting that axioms play a crucial role in constraining the interpretations of the language used to formalise the ontology. They define the intended models that correspond to the conceptualisation of the domain while excluding unintended interpretations. In this way, axioms help establish the boundaries and constraints of the ontology, ensuring its coherence and consistency. As an example, we can formulate simple axioms to define the properties of relations within our ontology. We can state that the relation *SupervisedBy* is asymmetric and intransitive, meaning that a junior researcher can be supervised by a professor, but not the other way around. On the other hand, the relation *CollaboratesWith* is symmetric, irreflexive, and non-transitive. These axioms help define the intended semantics of these relations within the domain. Several formal languages for representing ontologies exist, varying from schema languages (e.g., RDF schema) to more expressive logics (e.g., First Order and Higher-order logics, Modal Logic, Description Logics), to representation languages that go beyond formal semantics and make an explicit commitment to domain-independent foundational categories (e.g., OntoUML [22]). Clearly, expressive logical languages capture richer intended semantics, but do often not allow for sound and complete reasoning and if they do, reasoning sometimes remains untractable.

Description Logics (DL) [23] are one of the most well-known knowledge representation languages used to model ontologies in AI. They are of particular interest because they were created with the focus on tractable reasoning, and they provide the underpinning semantics of the W3C Web Ontology Language (OWL). Given a set of concept names, a set of role names, and some connectives that vary on the DL used, a DL ontology consists of two sets of axioms: the so-called TBox (terminological box) and the ABox (assertional box). In general, the TBox contains axioms describing relations between concepts, while the ABox contains axioms stating individuals (instances). For example, the statement 'every researcher is a person with a PhD title' belongs to the TBox, while 'Bob is a researcher' belongs to the ABox. Figure 1 shows a DL specification that we will consider here as an ontology for our simple conceptualisation of the university domain $\mathcal{EL}$ [24].[1] Notice that the example shows a very simple formalisation of the university domain; different formalisations may indeed exist, with different levels of details.

Ontologies played a crucial role as enabling technologies in the development of the Semantic Web [26]. The Semantic Web aims to annotate data on the web with semantic information, enabling computers to interpret and process data effectively. Although the Semantic Web did not fully realise all of its envisioned potential, ontologies have regained increased popularity in recent years, partly due to the re-emergence of knowledge graphs. Notably, the introduction of Google's Knowledge Graph in 2012 contributed significantly to this resurgence.[2] Knowledge graphs, powered by ontologies, have proven to be valuable tools for organising and structuring vast amounts of data, enabling efficient data retrieval, knowledge discovery, and semantic reasoning. Reasoning over ontologies or knowledge graphs can be performed by means of standard knowledge representation formalisms (e.g., RDF, RDF Schema, OWL) and query languages (e.g., SPARQL, Cypher, Gremlin) just to name a few of them.

In recent years, a number of knowledge graphs have become available on the Web, offering valuable structured information. One prominent example is DBpedia [18], which constructs a knowledge graph by automatically extracting key-value pairs from Wikipedia infoboxes. These pairs are then mapped to the DBpedia ontology using crowdsourcing efforts. Another notable knowledge graph is ConceptNet [20], a freely available linguistic knowledge graph that integrates information from various sources such as WordNet, Wikipedia, DBpedia, and OpenCyc. These knowledge graphs provide valuable resources for semantic understanding, information retrieval, and knowledge discovery. For a more comprehensive description of knowledge graphs and related standards, we recommend referring to the works cited in [15, 27].

2.2. Ontologies and Explanations

The knowledge representation language used to formalise an ontology provides support for both standard and non-standard reasoning tasks. Standard reasoning tasks involve checking various properties, such as subsumption and satisfaction. Concept subsumption determines if the description of one concept, for example, *Researcher*, is more general than the description of another concept, like *Professor*. Concept satisfaction, on the other hand, determines if a concept can be instantiated by an individual. Other standard reasoning tasks, such as query answering, are not semantically distinct from subsumption and satisfaction. Non-standard reasoning tasks have emerged to address new challenges in knowledge-based systems. These tasks include concept abduction and explanation, concept similarity, concept rewriting, and unification, among others. These non-standard reasoning tasks sometimes extend beyond the traditional subsumption and satisfaction checks, offering more advanced capabilities for knowledge-based systems.

Given this background, we believe that ontologies can play a crucial role in the realm of explainability. In particular, their adoption can significantly contribute to the development of explanations in neuro-symbolic AI from various perspectives:

- **Reference Modelling**: Ontologies provide sound and consensual models. By utilising ontologies as reference models, it becomes possible to capture system requirements effectively and promote the reuse of components. Additionally, ontologies contribute to system accountability and facilitate knowledge sharing and reuse. The main challenges associated with this perspective lie in reaching a consensus in the community about what elements (and their granularity) need to be included in a reference model for explanations. The vast literature on explanation in philosophy can play role in this process [28].
- **Common-Sense Reasoning**: Ontologies serve as a foundation for enriching explanations with context-aware semantic information. They support common-sense reasoning, which is essential for effectively transmitting knowledge to users. This capability enhances the comprehensibility of explanations for human understanding. The main challenges associated with this perspective lie in aligning the data used in the statistical models with semantic knowledge.

[1] $\mathcal{EL}$ is a lightweight DL allowing only for conjunctions and existential restrictions. It supports polynomial reasoning and is widely used to specify biomedical ontologies, see e.g., [25].

[2] https://www.blog.google/products/search/introducing-knowledge-graph-things-not/, last accessed on 2023/07/31.

– **Knowledge Refinement and Complexity Management**: Ontologies enable the representation of abstraction and refinement, fundamental mechanisms underlying human reasoning. These mechanisms provide opportunities for integrating knowledge from diverse sources and customizing the specificity and generality levels of explanations based on specific user profiles or target audiences. Moreover, systematic approaches grounded in Ontology (capital 'O', i.e., as a philosophical discipline) can be used to reveal domain notions that are fundamental for explaining the propositional knowledge contained in an ontology. The main challenges associated with this perspective lie in adapting existing knowledge refinement and complexity management techniques to work in the context of explainability.

By leveraging ontologies in these ways, the development of explanations can be enhanced, ensuring a solid conceptual basis, facilitating understanding through common-sense reasoning, and enabling flexible knowledge abstraction and refinement.

In the following sections, we overview some of the existing approaches in the literature, and we position them according to these three proposed perspectives. The rational behind the selection of these approaches lies in their connection with ontologies and their usage w.r.t. the above perspectives. This article is not intended to be a comprehensive review of the state-of-the-art of ontologies in neuro-symbolic AI (see [29] for some advances regarding ontologies and neuro-symbolic AI).

3. Reference Modelling

Historically, ontologies served as explicit models for the conceptual development of information systems [30]. Task ontologies were created to represent generic problem-solving methods and facilitate the reuse of task-dependent knowledge across diverse domains and applications [31]. In this context, the Unified Problem-Solving Method description Language (UPML) served as a relevant resource for representing tasks and problem-solving methods as reusable, domain-independent components [30]. Another example is the Web Service Modeling Ontology (WSMO), which focused on describing different aspects associated with Semantic Web Services [32]. These ontology resources played crucial roles in facilitating the representation and standardisation of knowledge in their respective classes of applications.

In the context of explainability, ontologies can play a crucial role as a common reference model for specifying explainable systems, i.e., as a model that addresses the area of explainable systems itself as a domain whose shared conceptualisation needs to be articulated and explicitly represented. Several studies have explored this avenue (e.g., [33–37]) highlighting the necessity of a shared interchange model for addressing the factors involved in explainable systems. To achieve this, they proposed taxonomies and ontologies to model key notions of the XAI domain including: explanations, users, the mapping of end-user requirements to specific explanation types, as well as to the AI capabilities of systems.

Nunes and Jannach [34] conducted a systematic literature review that examined the characteristics of explanations provided to users. Their work focused on aspects such as content, presentation, generation, and evaluation of explanations. They proposed a taxonomy that encompasses various explanation goals and different forms of knowledge that comprise the explanation components.

Arrieta et al. [33] present a taxonomy that established a mapping between deep learning models and the explanations they generate. Furthermore, they identified the specific features within these models that are responsible for generating these explanations. Their research contributes to the understanding of the relationship between model characteristics and the interpretability of their outputs. Their taxonomy covers different types of explanations that are produced by sub-symbolic models, including *simplification*, *explanation by examples*, *local explanations*, *text explanations*, *visual explanations*, *feature relevance*, and *explanations by transparent models*.

In the study by Wang et al. [35], the authors introduced a conceptual framework that elucidates how human reasoning processes can be integrated with explainable AI techniques. This framework establishes connections between different facets of explainability, such as explanation goals and types, as well as human reasoning mechanisms and AI methods. Notably, it facilitates a deeper understanding of the parallels between human reasoning and the generation of explanations by AI systems. By leveraging this conceptual framework, researchers and practitioners can gain insights into the interplay between human cognition and explainable AI.

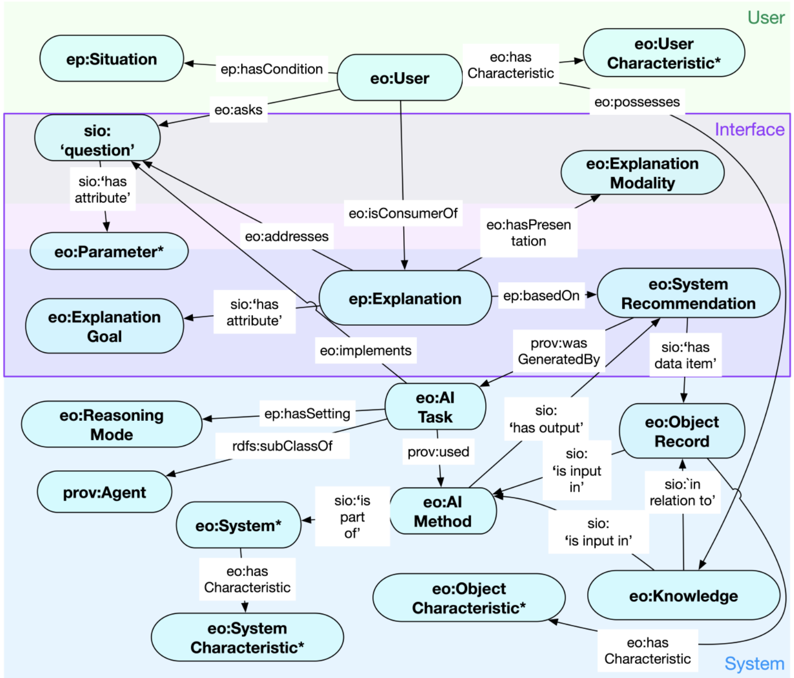

Fig. 2. Explanation ontology overview with key classes separated into three overlapping attribute categories: user, interface, and system [37].

Tiddi et al. [36] proposed an ontology design pattern specifically tailored for explanations. The authors observed that while the components of explanations may vary across different fields, there exist certain atomic components that can represent generic explanations. These atomic components include the associated *event*, the underlying *theory*, the *situation* to which the explanations are applied, and the *condition* that the explanations rely on. Additionally, the authors employ standard nomenclature such as *explanandum* (referring to what is being explained) and *explanans* (referring to what does the explaining) to further clarify the structure of explanations.

In a related study, Chari et al. [37] extended the ontology design pattern proposed by Tiddi et al. [36] to encompass explanations generated through computational processes. They developed an *explanation ontology* with the aim of facilitating user-centered explanations that enhance the explainability of model recommendations. This general-purpose model enables system designers to establish connections between explanations and the underlying data and knowledge. The explanation ontology incorporates attributes that form the foundation of explanations, such as *system recommendation* and *knowledge*, and models their interactions with the *questions* being addressed. It broadly categorises the dependencies of explanations into user, system, and interface attributes (see Figure 2). User attributes capture concepts related to a user consuming an explanation, including the user's question, characteristics, and situation. System attributes encompass concepts associated with the AI system used to generate recommendations and produce explanations. Explanations are defined as a taxonomy of literature-derived explanation types, with refined definitions of nine specific types, namely *case-based*, *contextual*, *counterfactual*, *everyday*, *scientific*, *simulation-based*, *statistical*, and *trace-based* explanations. These explanation types serve different user-centric purposes and vary in their suitability for different user *situations*, *context* and *knowledge* levels. They are generated by various *AI Task* and *methods*, each with distinct informational requirements. Interface attributes capture the intersection between user and system attributes that can be directly interacted with on a user interface. The explanation ontology developed by Chari et al. [37] is publicly available at https://tetherless-world.github.io/explanation-ontology.

Overall, ontologies act as a lingua franca for representation and information exchange in explainable AI systems by providing shared, formal representation of domain knowledge. They facilitate transparent communication, promote collaboration between different stakeholders, and enhance the interpretability and reliability of AI systems' explanations. By using ontologies, neuro-symbolic AI systems can bridge the gap between technical AI components and human understanding, making AI more transparent, accessible and, hence, trustworthy to end-users. However,

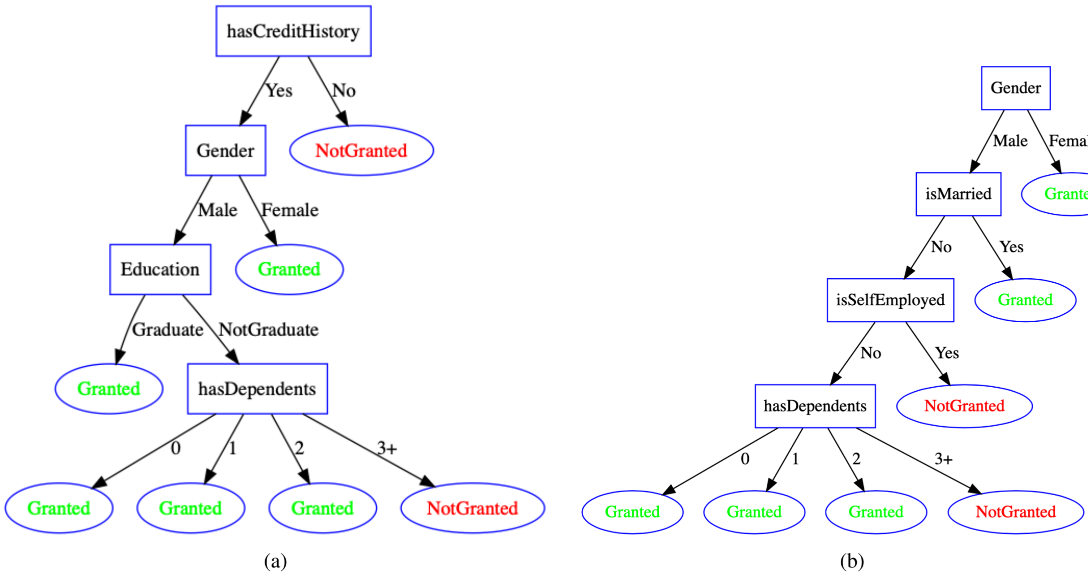

Fig. 3. Decision tree extracted without (a) and with (b) a domain ontology to explain the conditions to grant or refuse a loan [12]. It can be seen that the use of an ontology leads to different features appearing in the decision nodes.

despite the existence of several proposals for a general-purpose explanation ontology, they have primarily been utilized in an academic context. Their widespread adoption hinges on their standardization and acceptance within industry.

4. Common-sense Reasoning

The majority of existing approaches to XAI rely on statistical analysis of black box models [7]. While this type of analysis has demonstrated its utility in gaining some insight into the internal workings of black box models, it generally lacks support for explainability based on common-sense reasoning [8, 11]. As a result, these approaches often fall short in providing explanations that closely align with human reasoning, thereby limiting their capacity to (in the best scenario) generating intelligible explanations. Conversely, there is widespread acceptance that symbolic knowledge can effectively facilitate common-sense reasoning. Therefore, it is reasonable to consider that explanation techniques can leverage ontologies to enhance model explainability and generate explanations that are more understandable to humans.

In the context of explainability, several works attempt to fertilise explainability with ontologies. Seeliger et al. [38] surveyed what combinations of ontologies and knowledge graphs, and statistical models have been proposed to enhance model explainability, and what domains have been particularly important. The authors highlighted that quite a few approaches exist in supervised and unsupervised machine learning, whereas the integration of symbolic knowledge in reinforcement learning is almost overlooked. Most approaches in supervised learning seek to define a mapping between network inputs or neurons and ontology concepts which are then used in the explanations [12, 39, 40].

In general, these methods rely on the presence of a domain ontology, which aids in generating symbolic justifications for the outputs of neural network models. Nevertheless, the way in which the ontology is integrated may differ among various approaches. In [39], it is shown how the activations of the inner layers of a neural network w.r.t. a given sample can be aligned with domain ontology concepts. To detect this alignment, multiple mapping networks (one network for each concept) are trained. Each network takes certain layer activations as input and produces the relevance probability for the corresponding concept as output. However, since the number of inner activations is typically substantial, and not all concepts can be extracted from each layer, this alignment procedure can be inefficient. In [40], it is assumed that training data contains labels that are (manually) mapped to concepts defined in a domain ontology. This semantic link is then exploited to provide explanations as justifications of the classification obtained. In [12], a similar mapping is applied between input features and concepts in a domain ontology, where the ontology is used to guide the search for explanations. In particular, the authors proposed an algorithm that extracts decision trees as surrogate models of a black box classifier and takes into account ontologies in the tree extraction procedure. The algorithm learns decision trees whose decision nodes are associated with more general concepts defined in an ontology (Figure 3). This has proven to enhance the human-understandability of decision trees [12].

In the context of unsupervised learning, Tiddi et al. [41] introduced a technique to explain clusters by traversing a knowledge graph in order to identify commonalities among them. The system generates potential explanations by utilising both the background knowledge and the given cluster, and it is independent of the specific clustering algorithm employed.

One of the main challenges in all of these approaches lies in aligning the data utilised in statistical models with semantic knowledge. One possible solution is to create an ontology dedicated to each dataset and application, ensuring that the generated explanations are tailored to the specific problem. However, this approach can be prohibitive to application scenarios with stringent time and scalability constraints. An alternative in these cases might be to systematically construct a suitable ontology for mapping. This involves mapping sets of features to existing general domain ontologies, such as MS Concept Graph [19] or DBpedia [18]. This process can be facilitated through ontology matching techniques [42] and mapping methods [43]. It is important to note that human supervision is required for this process, and manual fine-tuning of the mappings is necessary. Unfortunately, there is no definitive blueprint to follow in this regard. The inclusion of explicit knowledge is essential for any attempts at generating human-interpretable explanations. The choice between a domain-specific ontology or adapting a domain-independent (i.e., upper-level, foundational) one on an ad-hoc basis depends on the specific requirements of each application. For example, when a primary requirement of the application is knowledge sharing and interoperability, opting for an upper-level/foundational ontology is strongly advised. Such an ontology can properly support the tasks of: making explicit the domain conceptualization at hand; safely identifying the correct relations between elements in different applications; reusing standardized domain-independent concepts and theories.

There are ontology matching approaches in the literature that establish mappings between domain ontologies with support of foundational ontologies, as well as approaches that map domain ontologies to foundational ontologies. In general, mapping knowledge derived from statistical techniques to ontologies (as computational logical theories) allows for supporting a form of hybrid reasoning, in which symbolic automated reasoning can complement statistical ones and circumvent their limitations [44]. However, in the particular case of mapping learned concepts and relations to a foundational ontology, we have the additional opportunity of grounding these knowledge elements in domain-independent axiomatized theories that describe common-sensical notions such as parthood, (existential, generic, historical) dependence, causality, temporal ordering of events, etc. [22]. This can, in turn, support more refined and transparent explanations via the expansion of the consequences of these mappings enabled by logical reasoning. For example, by mapping an element E to an ontological notion of event in a *event ontology* [45], one could infer a number of additional things about E (e.g., that it unfolds in time, that it has at least one object as participant, that it has a spatiotemporal extension, that its participants exist during the time the event is occurring, that it is a manifestation of some disposition, etc.).

5. Knowledge Refinement and Complexity Management

Abstraction and refinement are mechanisms that can be availed to represent knowledge in a more general or more specific manner. With abstraction one would hope to consider all is relevant and drop all the irrelevant details, with refinement one would hope to get more details. These mechanisms play a central role in human-reasoning. Humans use abstraction as a way to deal with the complexity of the world every day. Abstraction and refinement have many important applications, e.g., in natural language understanding, problem solving and planning, and reasoning by analogy.

Different formalisation of abstraction and refinement were proposed in the literature. Keet [46] argued that most proposals for abstractions differ along three dimensions: language to which it is applied, methodology, and semantics of what one does when abstracting. A syntactic theory of abstraction, mainly based on proof-theoretic notions, was

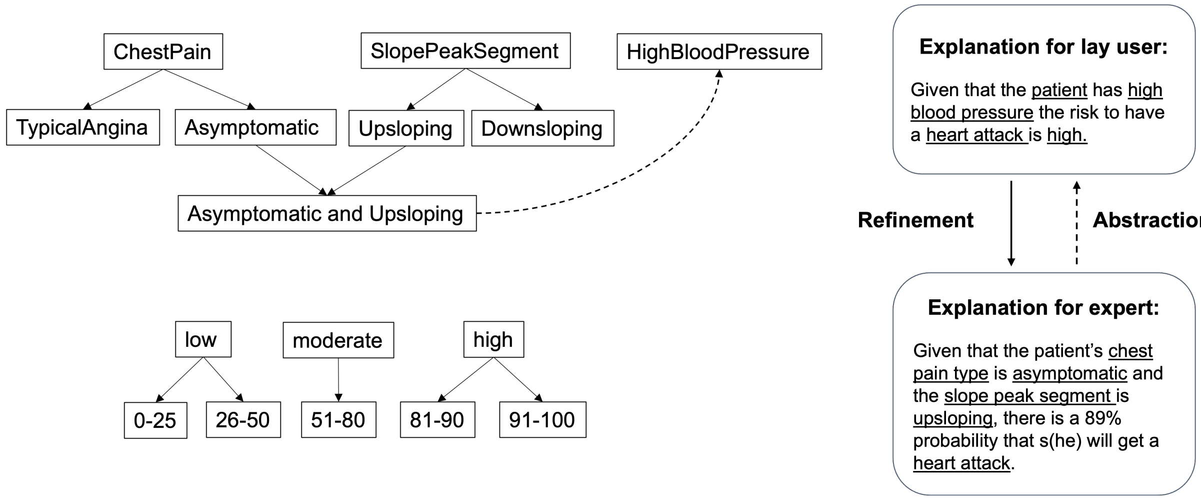

Fig. 4. Refinement and abstraction applied to obtain explanations tailored to expert and lay users. Explanations are grounded to a domain ontology modeling concept definitions and relations between them. An explanation can be made more specific or more general by exploiting concept relationships defined in the ontology. In the explanation for expert users, the concepts associated with *asymptomatic chest pain* and *upsloping peak segment* can be replaced by the more general concept of *high blood pressure*. Similarly, the predicted *probability value 89%* of getting a *heart attack* can be replaced by its qualitative description *high*. In this way, it is is possible to obtain an explanation that abstracts from technical details and that can be more suitable for lay users.

proposed in [47], whilst a semantic theory of abstractions based on viewing abstractions as model level mappings was proposed in [48]. These solutions were mainly theoretical and not developed for or assessed on their potential for implementation, reusability and scalability.

Abstraction is tightly connected with analogical reasoning. The structure mapping theory proposed by Gentner [49] suggests that humans use analogical reasoning to map the structure of one domain onto another, highlighting the shared relational information between objects. This process involves abstraction, as it can disregard specific object details in favor of identifying higher-level relational patterns that are relevant for making inferences and solving problems.

In the context of explainababilty, abstraction and analogical reasoning are essential for generating intelligible explanations to users. Abstraction allows an explanatory system to simplify complex decision-making processes by identifying higher-level patterns or by composing local explanations into (more general) global explanations [50]. On the other hand, analogical reasoning aids in aligning these patterns with familiar analogs from known domains. Consequently, users can grasp intricate decision-making processes in a more accessible and intelligible manner.

Abstraction and analogical reasoning are closely linked to the idea that explanations are selective and can be seen as truthful approximations of reality [11]. One typical example is the scientific explanation of an atom through the analogy of a miniature solar system, where the nucleus plays the role of the sun and the electrons orbit around it like planets. Clearly, this analogy is a simplified representation of the atomic structure. In reality, atoms are much more complex, and the behavior of electrons is better described using quantum mechanics. However, by using the solar system analogy, lay users can visualise and understand the basic concept of how an atom structure is organised with a central nucleus and orbiting electrons. Furthermore, explanations often involve simplifications to make the information more understandable. These simplifications are necessary because the complete representation of reality can be difficult to grasp. For instance, if asked to explain a prediction, a medical diagnosis AI system can offer an abstracted explanation that emphasizes the essential factors contributing to the diagnosis, rather than overwhelming the user with an exhaustive list of all the features and their values.

To exemplify this idea further, let us imagine that the medical diagnosis AI system runs a predictive model about the risk of a patient of getting a heart attack, and that it needs to provide an explanation for each prediction. The recipient of the explanation can be a medical doctor, who is a domain expert, or a lay user (Figure 4). On the one hand, the doctor would like to get a detailed explanation that provides insights about what attributes are used to make a diagnosis. On the other hand, a lay user might feel more comfortable in receiving a more abstract explanation that can still justify the diagnosis. Figure 4 exemplifies this idea. On the right side of the figure, two explanations are shown: one for an expert user and one for a lay user. Clearly, both of them explain the diagnosis but with a different level of details.[3] The explanations are grounded to a domain ontology, e.g., some features in the explanation are associated with concepts defined in the domain ontology. On the left side of the figure an excerpt of a simple ontology modeling the heart disease domain is represented. The ontology consists of simple concepts such as *ChestPain*, *Asymptomatic*, *TypicalAngina*, *SlopePeakSegment*, *Downsloping*, *Upsloping*, *HighBloodPressure*, a mapping between predicted probabilities of heart attacks to qualitative descriptions (e.g., a probability value in the range $[0, 50]$ is mapped to *low*-risk), and relationships between them. Some concepts are more specific than others, for instance, *Asymptomatic* and *TypicalAngina* are types of *ChestPain*. Simple concepts can be used to define complex ones, for instance, the concept built by the conjunction of *Asymptomatic* and *Upsloping* defines *HighBloodPressure*. Once the explanations are grounded to the domain ontology, these can be made more specific or more abstract by exploiting the relationships defined in the ontology. For instance, the concepts related to chest pain type *Asymptomatic* and peak segment *Upsloping* can be replaced by *HighBloodPressure*. Similarly, the 89% probability of getting a heart attack can be replaced by its qualitative description *high*-risk. In this way, it is is possible to obtain an explanation that abstracts from technical details and that can be more suitable for lay users.

Abstraction is just one of the possible techniques to address the goals of complexity management of complex information structures [54]. Complexity Management, more generally, is essential for explanation, given that to explain one must focus on the explanation goals of an explanation-seeker [55]. In other works, according to the *Pragmatic Approach to Explanation* [56], to explain is to answer the *requests-for-explanation* (*why-questions*) of an explanation-seeker. To do that, one has to dispense with information that does not contribute to that goal, i.e., to declutter the information structures and processes at hand. One can, for example, leverage the grounding of domain ontologies in foundational ones to automatically perform complexity management operations over large domain models. These operations include beside abstraction (ontology summarisation), modularisation and viewpoint extraction [54, 57].

Another important notion of refinement is *ontological unpacking* [28], a type of explanation (as well as type of knowledge refinement) that aims at revealing the ontological semantics underlying a given symbolic artifact. Ontological unpacking is a methodology that operationalises the notions of *truthmaking* and grounding in metaphysical explanation but also incorporating elements from the *unificatory approach* [58], namely the use of *foundational ontology design patterns* as an explanatory device. Design Patterns are the modeling analog of metaphors, i.e., they represent exactly the structure that is preserved across different (albeit analogous) situations. As discussed in [28], they can play a role analogous to the notion of a *schematic structure/argument patterns* in the unificatory approach. In the latter, these patterns support explanation as *theoretical reduction*, i.e., reducing (as in structural transferring [49] or analogical reasoning) relevant aspects of a phenomenon to be explained to those of an understood phenomenon. In other words, theoretical reduction and unification using these patterns provide a mechanism for explanation employing analogy but also a mechanism for complexity management, since it puts focus on the elements that one needs to focus on, namely, exactly those constituting the pattern at hand. Ontological Unpacking has been shown to be effective in generating semantically transparent and unambiguous human-understandable explanations in complex domains [59, 60]. It can be noted that ontology unpacking (as all refinement operations) goes in the inverse direction of abstraction [55]. Thus, ontology unpacking must be complemented by the aforementioned complexity management techniques.

Knowledge refinement and complexity management techniques play a pivotal role in offering explanations tailored to user requirements and backgrounds. Despite notable progress in their adoption, there remains scope for their advancement as enabling technologies for explainability in neuro-symbolic AI. For instance, ontology unpacking

[3] An explanation for a given instance can be obtained using a local explanation method such as LIME [51] or ANCHOR [52]. These methods explain the prediction of specific instances by identifying what attributes or features contribute more to the prediction. Once relevant attributes have been identified, a textual explanation can be generated by filling some explanation templates or by means of more advanced natural language processing techniques [53].

currently relies heavily on manual processes, necessitating collaboration with domain experts, while the development and exploration of supportive tools are ongoing endeavors. Abstraction and refinement tasks often deal with extensive search spaces, with the incorporation of preferences or heuristics still to be explored.

6. Open Challenges

Apart from the challenges associated with the three perspectives discussed until now, there are several other open challenges that must be tackled to integrate ontologies as fundamental enabling technologies for explanations in neuro-symbolic AI. In the following, we outline a few of them. The interested reader will find a more comprehensive discussion of the challenges associated with explanations in [8, 61].

- Ontology as explanations: While ontologies may contribute to offer explanations within a domain of interest (see e.g., [12, 62]), and, ontology usage should result in an understanding of the domain, this does not automatically mean that the ontology itself is self-explainable and easily understandable by humans [55]. The same holds for other symbolic artefacts that are offered as explanations to numerical black boxes (e.g., knowledge graphs, decisions trees). For this reason, it can be argued that these symbolic artefacts also require their own explanation. The idea of ontological unpacking is relevant here as a means to identify and make explicit the *truthmakers* of the propositions represented in that domain model, i.e., the entities in the world that make those propositions true, and from which the logico-linguistic constructions that constitute symbolic models are derived [28]. This, in turn, can enhance the human understandability of explanations.
- Causal explanations: A key notion related to explanations is that of causality [63]. Although not all explanations are causal explanations, causal explanations occupy a fundamental place in the scientific explanation literature [64]. In particular, knowing what relationship there is between input and output, or between input features can foster human-understandable explanations. However, causal explanations are largely lacking in the machine learning literature, with only few exceptions, e.g., [65]. Ontologies can capture causal rules once the knowledge over an application domain has been modelled. On the other hand, to the best of our knowledge, only a few works attempted to define and model causality within an ontology. In [66], the authors proposed different definitions of causality, and studied how constraints of different nature, namely structural, causality, and circumstantial, intervene in shaping causal relations. However, as the authors claim, the approach is incomplete and further extensions are needed. [45] puts forth a formal theory of causation. In this theory, we have the explicit manifestation of dispositions, which are activated by the obtaining of certain situations (roughly individual state of affairs), and which are manifested via the occurrence of events. In other words, events cause each other via a continuous mechanism of events bringing about situations, that activate dispositions, that in turn are manifested as other events, and so on.
- Evaluating the human-understandability and effectiveness of explanations: Rudin et al. [67] pointed out evaluation of explanations as a major challenge to be faced in the context of XAI. On the one hand, one would like to quantify to what extent users can understand and use an explanation. A few approaches proposed quantitative metrics and protocols [68–70], but it is still unclear how to compare the results of different evaluations and establish a common understanding of how to evaluate explanations. There are already some promising approaches in the literature to solve this problem. In [71], the authors identify several conceptual properties that should be considered to assess the quality of explanations, and they propose quantitative evaluation methods to evaluate an explanation. More recently, a survey-based methodology for guiding the human evaluation of explanations was proposed in [72].
- Ontologies and large language models: The capability of large language models (LLMs) to generate multimodal content, which can sometimes result from hallucinations, raises doubts about their trustworthiness. Several works attempt to connect ontologies with LLMs [73, 74]. Once a link between the content generated by an LLM and an ontology is established, the latter can be used to check the consistency of the content, provide explanations for its consistency or inconsistency, and potentially offer ways to repair hallucinations.

Finally, another ontology-based approach to explanation, which is connected to value-based justification (in ethics) is the one discussed in [9]. There, explicability is related to the reconstruction of decision-making processes, which in turn are grounded on preference relations, which in turn are grounded on value-assessments. The whole approach is grounded on an ontological analysis of ethical dimensions, and, ultimately, on an ontological analysis of the notions of value, risk, autonomy and delegation.

7. Summary and Conclusion

In the last years, there has been a resurgence of interest in symbolic AI. Symbolic AI stands out as a pivotal enabling technology for neuro-symbolic AI systems. It effectively addresses the constraints inherent in black box deep learning models by facilitating reasoning capabilities and explanatory support.

In this paper, we discussed the role of ontologies and knowledge in explanations for neuro-symbolic AI from three perspectives: reference modelling, common-sense reasoning, and knowledge refinement and complexity management.

The role played by ontologies within these perspectives can be summarized as follows. Firstly, ontologies provide formal reference consensual models for designing explainable systems and generating human-understandable explanations. Ontologies provide a common lingua for defining explanations, promoting interoperability, and reusability of explanations across various domains. Secondly, ontologies enable the creation of explanations with linked semantics. This can, in turn, support more refined and transparent explanations via knowledge expansion enabled by logical reasoning. Thus, integrating ontologies with current explanation techniques allows for supporting a form of hybrid reasoning, enhancing the human-understandability of explanations. Finally, ontologies offer the ability to abstract and refine knowledge, which serves as the foundation for human reasoning. Knowledge refinement and complexity management is essential to craft personalised explanations that are human-centric and tailored to different user profiles.

Given the above, ontologies can play a crucial role for explanations in neuro-symbolic AI. Nevertheless, a number of challenges still need to be addressed, namely the integration of foundational and domain ontologies in current explainability approaches, the adoption of complexity management techniques to ensure ontologies as explanations are easily manageable and comprehensible for users, how to model and evaluate causality, and the evaluation of their human-understandability. Last but not least, an important challenge to address is establishing the relationship between ontologies and large language models, and exploring how this connection can be used to explain and possibly correct LLMs hallucinations.

References

[1] A.d. Garcez and L.C. Lamb, Neurosymbolic AI: the 3rd wave, *Artificial Intelligence Review* (2023). doi:10.1007/s10462-023-10448-w.

[2] H. Kautz, The Third AI Summer: AAAI Robert S. Engelmore Memorial Lecture, *AI Magazine* **43**(1) (2022), 105–125. doi:10.1002/aaai.12036.

[3] Z. Zhang, L. Yilmaz and B. Liu, A Critical Review of Inductive Logic Programming Techniques for Explainable AI, *IEEE Transactions on Neural Networks and Learning Systems* (2023), 1–17. doi:10.1109/TNNLS.2023.3246980.

[4] S. Badreddine, A. d'Avila Garcez, L. Serafini and M. Spranger, Logic Tensor Networks, *Artificial Intelligence* **303** (2022), 103649. doi:https://doi.org/10.1016/j.artint.2021.103649.

[5] M. Richardson and P. Domingos, Markov Logic Networks, *Machine Learning* **62**(1–2) (2006), 107–136–. doi:10.1007/s10994-006-5833-1.

[6] R. Riegel, A.G. Gray, F.P.S. Luus, N. Khan, N. Makondo, I.Y. Akhalwaya, H. Qian, R. Fagin, F. Barahona, U. Sharma, S. Ikbal, H. Karanam, S. Neelam, A. Likhyani and S.K. Srivastava, Logical Neural Networks, *CoRR* **abs/2006.13155** (2020).

[7] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J.M. Alonso-Moral, R. Confalonieri, R. Guidotti, J.D. Ser, N. Díaz-Rodríguez and F. Herrera, Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence, *Information Fusion* (2023), 101805. doi:https://doi.org/10.1016/j.inffus.2023.101805.

[8] R. Confalonieri, L. Coba, B. Wagner and T.R. Besold, A historical perspective of explainable Artificial Intelligence, *WIREs Data Mining and Knowledge Discovery* **11**(1) (2021). doi:https://doi.org/10.1002/widm.1391.

[9] R. Guizzardi, G. Amaral, G. Guizzardi and J. Mylopoulos, An ontology-based approach to engineering ethicality requirements, *Software and Systems Modeling* (2023), 1–27.

[10] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti and D. Pedreschi, A Survey of Methods for Explaining Black Box Models, *ACM Comp. Surv.* **51**(5) (2018), 1–42.

[11] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, *Artificial Intelligence* **267** (2019), 1–38. doi:https://doi.org/10.1016/j.artint.2018.07.007.

[12] R. Confalonieri, T. Weyde, T.R. Besold and F.M. del Prado Martín, Using ontologies to enhance human understandability of global post-hoc explanations of Black-box models, *Artificial Intelligence* **296** (2021). doi:https://doi.org/10.1016/j.artint.2021.103471.

[13] B. Mittelstadt, C. Russell and S. Wachter, Explaining Explanations in AI, in: *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, ACM Press, New York, New York, USA, 2019, pp. 279–288. ISBN 9781450361255. doi:10.1145/3287560.3287574.

[14] S. Chari, D.M. Gruen, O. Seneviratne and D.L. McGuinness, Directions for Explainable Knowledge-Enabled Systems, in: *Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges*, I. Tiddi, F. Lécué and P. Hitzler, eds, Studies on the Semantic Web, Vol. 47, IOS Press, 2020, pp. 245–261. doi:10.3233/SSW200022.

[15] I. Tiddi and S. Schlobach, Knowledge graphs as tools for explainable machine learning: A survey, *Artificial Intelligence* **302** (2022), 103627. doi:https://doi.org/10.1016/j.artint.2021.103627.

[16] A. Iana, M. Alam and H. Paulheim, A survey on knowledge-aware news recommender systems, *Semantic Web Journal* (2022).

[17] P. Ristoski and H. Paulheim, Semantic Web in data mining and knowledge discovery: A comprehensive survey, *Journal of Web Semantics* **36** (2016), 1–22. doi:10.1016/j.websem.2016.01.001.

[18] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P.N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer and C. Bizer, DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia, *Semantic Web Journal* **6**(2) (2015), 167–195.

[19] W. Wu, H. Li, H. Wang and K.Q. Zhu, Probase: A Probabilistic Taxonomy for Text Understanding, in: *Proc. of the 2012 ACM SIGMOD Int. Conf. on Management of Data*, SIGMOD '12, ACM, 2012, pp. 481–492. ISBN 978-1-4503-1247-9.

[20] R. Speer, J. Chin and C. Havasi, ConceptNet 5.5: An Open Multilingual Graph of General Knowledge, in: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, AAAI Press, 2017, pp. 4444–4451–.

[21] N. Guarino, D. Oberle and S. Staab, What Is an Ontology?, in: *Handbook on Ontologies*, S. Staab and R. Studer, eds, International Handbooks on Information Systems, Springer, 2009, pp. 1–17. ISBN 978-3-540-92673-3.

[22] G. Guizzardi, A. Botti Benevides, C.M. Fonseca, D. Porello, J.P.A. Almeida and T. Prince Sales, UFO: Unified foundational ontology, *Applied Ontology* **17**(1) (2022), 167–210.

[23] F. Baader, D. Calvanese, D.L. McGuinness, D. Nardi and P.F. Patel-Schneider (eds), *The Description Logic Handbook: Theory, Implementation, and Applications*, Cambridge University Press, New York, NY, USA, 2003. ISBN 0-521-78176-0.

[24] F. Baader, S. Brandt and C. Lutz, Pushing the $\mathcal{EL}$ Envelope., in: *IJCAI*, 2005, pp. 364–369.

[25] N. Matentzoglu and B. Parsia, BioPortal Snapshot 30.03.2017, 2017, last accessed, 2022/08/11. doi:10.5281/zenodo.439510.

[26] T. Berners-Lee, J. Hendler and O. Lassila, The Semantic Web, *Scientific American* **284**(5) (2001), 34–43.

[27] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutiérrez, J.E.L. Gayo, S. Kirrane, S. Neumaier, A. Polleres, R. Navigli, A.N. Ngomo, S.M. Rashid, A. Rula, L. Schmelzeisen, J.F. Sequeda, S. Staab and A. Zimmermann, *Knowledge Graphs*, Springer, 2021.

[28] G. Guizzardi and N. Guarino, Semantics, Ontology and Explanation, *CoRR* **abs/2304.11124** (2023).

[29] P. Hitzler, Some advances regarding ontologies and neuro-symbolic artificial intelligence, in: *ECMLPKDD Workshop on Meta-Knowledge Transfer*, P. Brazdil, J.N. van Rijn, H. Gouk and F. Mohr, eds, Proceedings of Machine Learning Research, Vol. 191, PMLR, 2022, pp. 8–10. https://proceedings.mlr.press/v191/hitzler22a.html.

[30] D. Fensel, E. Motta, F. van Harmelen, V.R. Benjamins, M. Crubezy, S. Decker, M. Gaspari, R. Groenboom, W. Grosso, M. Musen, E. Plaza, G. Schreiber, R. Studer and B. Wielinga, The Unified Problem-Solving Method Development Language UPML, in: *Knowledge and Information Systems*, Vol. 5, 2003, pp. 83–131. ISSN 0219-1377. ISBN 1011500200745. doi:10.1007/s10115-002-0074-5.

[31] N. Guarino, *Formal ontology in information systems: Proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy*, Vol. 46, IOS press, 1998.

[32] D. Fensel and C. Bussler, The web service modeling framework WSMF, *Electronic Commerce Research and Applications* **1**(2) (2002), 113–137. doi:10.1016/S1567-4223(02)00015-7.

[33] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila and F. Herrera, Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Information Fusion* **58**(October 2019) (2020), 82–115. doi:10.1016/j.inffus.2019.12.012.

[34] I. Nunes and D. Jannach, A systematic review and taxonomy of explanations in decision support and recommender systems, *User Modeling and User-Adapted Interaction* **27**(3–5) (2017), 393–444. doi:10.1007/s11257-017-9195-0.

[35] D. Wang, Q. Yang, A. Abdul and B.Y. Lim, *Designing Theory-Driven User-Centric Explainable AI*, in: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–15–. ISBN 9781450359702.

[36] I. Tiddi, M. D'Aquin and E. Motta, An ontology design pattern to define explanations, *Proceedings of the 8th International Conference on Knowledge Capture, K-CAP 2015* (2015). ISBN 9781450338493. doi:10.1145/2815833.2815844.

[37] S. Chari, O. Seneviratne, M. Ghalwash, S. Shirai, D.M. Gruen, P. Meyer, P. Chakraborty and D.L. McGuinness, Explanation Ontology: A general-purpose, semantic representation for supporting user-centered explanations, *Semantic Web* **Preprint** (2023), 1–31, Preprint. doi:10.3233/SW-233282.

[38] A. Seeliger, M. Pfaff and H. Krcmar, Semantic web technologies for explainable machine learning models: A literature review, in: *Joint Proceedings of the 6th International Workshop on Dataset PROFILing and Search & the 1st Workshop on Semantic Explainability co-located with the 18th International Semantic Web Conference (ISWC 2019)*, CEUR Workshop Proceedings, Vol. 2465, 2019, pp. 30–45.

[39] M. de Sousa Ribeiro and J. Leite, Aligning Artificial Neural Networks and Ontologies towards Explainable AI, *Proceedings of the AAAI Conference on Artificial Intelligence* **35**(6) (2021), 4932–4940. doi:10.1609/aaai.v35i6.16626.

[40] Z. Daniels, L. Frank, C. Menart, M. Raymer and P. Hitzler, A framework for explainable deep neural models using external knowledge graphs, in: *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II*, 2020, p. 73. doi:10.1117/12.2558083.

[41] I. Tiddi, M. d'Aquin and E. Motta, Dedalo: Looking for Clusters Explanations in a Labyrinth of Linked Data, in: *The Semantic Web: Trends and Challenges*, Springer International Publishing, Cham, 2014, pp. 333–348. ISBN 978-3-319-07443-6.

[42] F. Giunchiglia and P. Shvaiko, Semantic matching, *Knowl. Eng. Rev.* **18**(3) (2003), 265–280. doi:10.1017/S0269888904000074.

[43] Y. Kalfoglou and W.M. Schorlemmer, Ontology mapping: the state of the art, *Knowl. Eng. Rev.* **18**(1) (2003), 1–31. doi:10.1017/S0269888903000651.

[44] Y. Li, S. Ouyang and Y. Zhang, Combining deep learning and ontology reasoning for remote sensing image semantic segmentation, *Knowledge-Based Systems* **243** (2022), 108469.

[45] A. Botti Benevides, J.-R. Bourguet, G. Guizzardi, R. Peñaloza and J.P.A. Almeida, Representing a reference foundational ontology of events in SROIQ, *Applied Ontology* **14**(3) (2019), 293–334.

[46] C.M. Keet, Enhancing Comprehension of Ontologies and Conceptual Models Through Abstractions, in: *Proc. of the 10th Congress of the Italian Association for Art. Intel. (AI*IA 2007)*, 2007, pp. 813–821.

[47] F. Giunchiglia and T. Walsh, A theory of abstraction, *Artificial Intelligence* **57**(2) (1992), 323–389. doi:https://doi.org/10.1016/0004-3702(92)90021-O.

[48] P.P. Nayak and A.Y. Levy, A Semantic Theory of Abstractions, in: *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95, Montréal Québec, Canada, August 20-25 1995, 2 Volumes*, Morgan Kaufmann, 1995, pp. 196–203.

[49] D. Gentner, Structure-mapping: A theoretical framework for analogy, *Cognitive Science* **7**(2) (1983), 155–170. doi:https://doi.org/10.1016/S0364-0213(83)80009-3.

[50] M. Setzu, R. Guidotti, A. Monreale, F. Turini, D. Pedreschi and F. Giannotti, GLocalX - From Local to Global Explanations of Black Box AI Models, *Artificial Intelligence* **294** (2021), 103457. doi:https://doi.org/10.1016/j.artint.2021.103457.

[51] M.T. Ribeiro, S. Singh and C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in: *Proc. of the 22nd Int. Conf. on Knowledge Discovery and Data Mining*, KDD '16, ACM, 2016, pp. 1135–1144. ISBN 978-1-4503-4232-2.

[52] M.T. Ribeiro, S. Singh and C. Guestrin, Anchors: High-Precision Model-Agnostic Explanations, in: *AAAI*, AAAI Press, 2018, pp. 1527–1535.

[53] N.F. Rajani, B. McCann, C. Xiong and R. Socher, Explain Yourself! Leveraging Language Models for Commonsense Reasoning, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4932–4942. doi:10.18653/v1/P19-1487.

[54] E. Romanenko, D. Calvanese and G. Guizzardi, Abstracting Ontology-Driven Conceptual Models: Objects, Aspects, Events, and Their Parts, in: *Research Challenges in Information Science*, R. Guizzardi, J. Ralyté and X. Franch, eds, Springer International Publishing, Cham, 2022, pp. 372–388. ISBN 978-3-031-05760-1.

[55] E. Romanenko, D. Calvanese and G. Guizzardi, Towards Pragmatic Explanations for Domain Ontologies, in: *Knowledge Engineering and Knowledge Management - 23rd International Conference, EKAW 2022, Bolzano, Italy, September 26-29, 2022, Proceedings*, Ó. Corcho, L. Hollink, O. Kutz, N. Troquard and F.J. Ekaputra, eds, Lecture Notes in Computer Science, Vol. 13514, Springer, 2022, pp. 201–208. doi:10.1007/978-3-031-17105-5_15.

[56] B.C. Van Fraassen, The pragmatics of explanation, *American philosophical quarterly* **14**(2) (1977), 143–150.

[57] Guizzardi, Giancarlo and Sales, Tiago Prince and Almeida, Joao Paulo A. and Poels, Geert, Automated conceptual model clustering : a relator-centric approach, *Software and Systems Modeling* **21** (2022), 1363–1387.

[58] P. Kitcher, Explanatory unification, *Philosophy of science* **48**(4) (1981), 507–531.

[59] A. Bernasconi, G. Guizzardi, O. Pastor and V.C. Storey, Semantic interoperability: ontological unpacking of a viral conceptual model, *BMC bioinformatics* **23**(11) (2022), 1–23.

[60] A. García, A. Bernasconi, G. Guizzardi, O. Pastor, V.C. Storey and I. Panach, Assessing the value of ontologically unpacking a conceptual model for human genomics, *Information Systems* (2023), 102242.

[61] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J.D. Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, R. Jiang, H. Khosravi, F. Lecue, G. Malgieri, A. Páez, W. Samek, J. Schneider, T. Speith and S. Stumpf, Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions, *Information Fusion* **106** (2024), 102301. doi:https://doi.org/10.1016/j.inffus.2024.102301. https://www.sciencedirect.com/science/article/pii/S1566253524000794.

[62] Z. Horne, M. Muradoglu and A. Cimpian, Explanation as a Cognitive Process, *Trends in Cognitive Sciences* **23**(3) (2019), 187–199, Publisher Copyright: © 2018 Elsevier Ltd. doi:10.1016/j.tics.2018.12.004.

[63] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd edn, Cambridge University Press, USA, 2009. ISBN 052189560X.

[64] P. Kitcher, Explanatory Unification and the Causal Structure of the World, in: *Scientific Explanation*, P. Kitcher and W. Salmon, eds, Minneapolis: University of Minnesota Press, 1989, pp. 410–505.

[65] A. Chattopadhyay, P. Manupriya, A. Sarkar and V.N. Balasubramanian, Neural Network Attributions: A Causal Perspective, in: *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, eds, Proceedings of Machine Learning Research, Vol. 97, PMLR, Long Beach, California, USA, 2019, pp. 981–990.

[66] J. Lehmann, S. Borgo, C. Masolo and A. Gangemi, Causality and Causation in DOLCE, in: *Proceedings of the international conference on Formal ontology in information systems (FOIS 2004)*, IOS Press, 2004, pp. 273–284.

[67] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova and C. Zhong, Interpretable machine learning: Fundamental principles and 10 grand challenges, *Statistics Surveys* **16**(none) (2022), 1–85. doi:10.1214/21-SS133.
[68] R.R. Hoffman, T. Miller, S.T. Mueller, G. Klein and W.J. Clancey, Explaining Explanation, Part 4: A Deep Dive on Deep Nets, *IEEE Intell. Systems* **33**(3) (2018), 87–95.
[69] G. Vilone and L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, *Information Fusion* **76** (2021), 89–106. doi:https://doi.org/10.1016/j.inffus.2021.05.009.
[70] M. Chromik and M. Schuessler, A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI, in: *ExSS-ATEC@IUI*, 2020.
[71] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen and C. Seifert, From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI, *ACM Comput. Surv.* **55**(13s) (2023). doi:10.1145/3583558.
[72] R. Confalonieri and J.M. Alonso-Moral, An operational framework for guiding human evaluation in Explainable and Trustworthy AI, *IEEE Intelligent Systems* (2024), 1–13. doi:10.1109/MIS.2023.3334639.
[73] B. Bi, C. Wu, M. Yan, W. Wang, J. Xia and C. Li, Incorporating External Knowledge into Machine Reading for Generative Question Answering, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng and X. Wan, eds, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 2521–2530. doi:10.18653/v1/D19-1255. https://aclanthology.org/D19-1255.
[74] A. Martino, M. Iannelli and C. Truong, Knowledge Injection to Counter Large Language Model (LLM) Hallucination, in: *The Semantic Web: ESWC 2023 Satellite Events*, C. Pesquita, H. Skaf-Molli, V. Efthymiou, S. Kirrane, A. Ngonga, D. Collarana, R. Cerqueira, M. Alam, C. Trojahn and S. Hertling, eds, Springer Nature Switzerland, Cham, 2023, pp. 182–185. ISBN 978-3-031-43458-7.