information

MDPI

Review

On the Integration of Knowledge Graphs into Deep Learning Models for a More Comprehensible AI—Three Challenges for Future Research

Giuseppe Futia * and Antonio Vetrò

Nexa Center for Internet and Society, Department of Control and Computer Engineering (DAUIN), Politecnico di Torino, 10129 Turin, Italy; antonio.vetro@polito.it

* Correspondence: giuseppe.futia@polito.it; Tel.: +39-0110907216

Received: 24 December 2019; Accepted: 18 February 2020; Published: 22 February 2020

Abstract: Deep learning models contributed to reaching unprecedented results in prediction and classification tasks of Artificial Intelligence (AI) systems. However, alongside this notable progress, they do not provide human-understandable insights on how a specific result was achieved. In contexts where the impact of AI on human life is relevant (e.g., recruitment tools, medical diagnoses, etc.), explainability is not only a desirable property, but it is -or, in some cases, it will be soon-a legal requirement. Most of the available approaches to implement eXplainable Artificial Intelligence (XAI) focus on technical solutions usable only by experts able to manipulate the recursive mathematical functions in deep learning algorithms. A complementary approach is represented by symbolic AI, where symbols are elements of a lingua franca between humans and deep learning. In this context, Knowledge Graphs (KGs) and their underlying semantic technologies are the modern implementation of symbolic AI—while being less flexible and robust to noise compared to deep learning models, KGs are natively developed to be explainable. In this paper, we review the main XAI approaches existing in the literature, underlying their strengths and limitations, and we propose neural-symbolic integration as a cornerstone to design an AI which is closer to non-insiders comprehension. Within such a general direction, we identify three specific challenges for future research—*knowledge matching*, *cross-disciplinary explanations* and *interactive explanations*.

Keywords: eXplainable artificial intelligence; deep learning; knowledge graphs

1. Introduction

Deep Learning techniques are dominant in the modern approach to Artificial Intelligence (AI). Their use is widespread, due to their very high performance in prediction and classification tasks across application areas [1–3]. However, alongside their large adoption, current deep learning models are opaque and do not provide human-understandable insights on their outputs. Explainability is a crucial requirement [4] if deep learning is used for tasks that heavily impact the lives of people, such as recruitment tools [5], decision support systems for justice [6], prevention of terrorism [7], clinical applications [8], just to mention a few sensitive domains.

EXplainable Artificial Intelligence (XAI) is the field of research where mathematicians, computer scientists and software engineers design, develop and test techniques for making AI systems more transparent and comprehensible by its stakeholders. Most of the approaches developed in this field require very specific technical expertise to manipulate the recursive algorithms that implement the mathematical functions at the roots of deep learning. Moreover, understanding this mathematical scaffolding is not enough to get insights into internal working models. In fact, in order to be more

understandable, deep-learning-based systems should be able to emit and manipulate symbols, enabling user explanations on how a specific result is achieved [4].

In the context of symbolic systems, Knowledge Graphs (KGs) [9] and their underlying semantic technologies are a promising solution for the issue of understandability [10]. In fact, these large networks of semantic entities and relationships provide a useful backbone for several reasoning mechanisms, ranging from consistency checking [11], to causal inference [12]. These reasoning procedures are enabled by ontologies [13], which provide a formal representation of semantic entities and relationships relevant to a specific sphere of knowledge.

We consider the implementations of symbolic systems based on semantic technologies suitable to improve explanations for non-insiders. Input features, hidden layers and computational units, and predicted output of deep learning models can be mapped into entities of KGs or concepts and relationships of ontologies (*knowledge matching*). Traditionally, these ontology artifacts are the results of conceptualizations and practices adopted by experts from various disciplines, such as biology [14], finance [15], and law [16]. As a consequence, they are very comprehensible to people with expertise in a specific domain (*cross-disciplinary explanations*), even if they do not have skills in AI technologies. Moreover, in the context of semantic technologies, KGs and ontologies are natively built to be queried and therefore they are able to provide answers to user requests (*interactive explanations*) and to provide a symbolic level to interpret the behaviour and the results of a deep learning model.

We represent a schematic graph that summarises the role of semantic technologies for XAI in Figure 1. In particular, knowledge matching of deep learning components, including input features, hidden unit and layers, and output predictions with KGs and ontology components can make the internal functioning of algorithms more transparent and comprehensible; in addition, query and reasoning mechanisms of KGs and ontologies enable the conditions for advanced explanations, namely cross-disciplinary and interactive explanations.

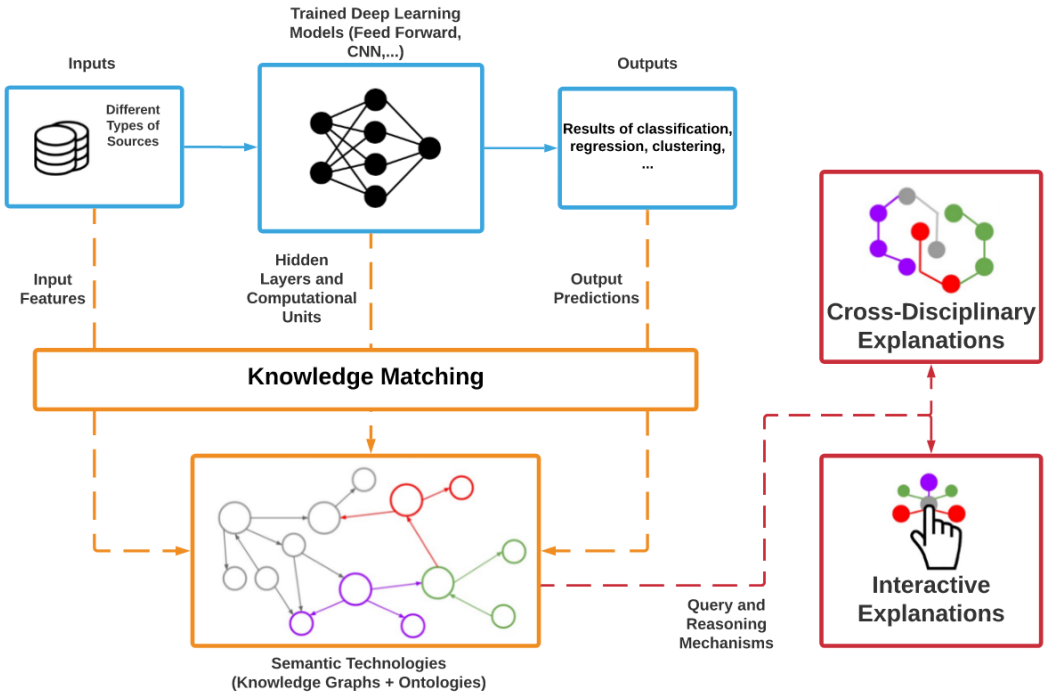

Figure 1. Schematic representation of an explainable artificial intelligence (AI) system that integrates semantic technologies into deep learning models. The traditional pipeline of an AI system is depicted with the blue color. The Knowledge Matching process of deep learning components with Knowledge Graphs (KGs) and ontologies is depicted with orange color. Cross-Disciplinary and Interactive Explanations enabled by query and reasoning mechanisms are depicted with the red color.

In this position paper our goal is to review the main XAI approaches existing in the literature, underlying their strengths and limitations, and propose neural-symbolic integration as a cornerstone to design an AI which is closer to non-insiders comprehension. In Section 2 we provide a background context in terms of learning approach between deep learning systems (connectionist AI), and reasoning systems built on top of ontologies and KGs (symbolic AI). Then, in Section 3 we outline the technical issues related to XAI, describing current solutions that require expertise on deep learning techniques. In Section 4 we provide details on three promising research challenges (knowledge matching, cross-disciplinary explanations, and interactive explanations) enabled by the integration of KGs and ontologies into deep learning models. We conclude with Section 5, where we synthesize our contribution and future research we identify for eXplainable AI.

2. Background: Learning Approach in Connectionist and Symbolic AI

In this Section we describe the main learning principles behind deep learning models (connectionist AI) and KGs and ontologies (symbolic AI). Understanding their differences is a crucial requirement to realize how symbolic systems can be integrated within connectionist approaches to build a more comprehensible AI.

The movement of *connectionism*, also known as parallel distributed processing, correspond to "the second wave" (the first wave dates back to the 1940s and it was based on a neural perspective of AI [17]) of research in neural networks. Born in the 1980s, the central idea of connectionists is that a network of simple computational units, defined as *neurons*, excite and inhibit each other in parallel to achieve an intelligent behavior. As underlined by Rumelhart et al. [18], the resulting knowledge consists of the connections between these computational units distributed throughout the network. Two key concepts of the connectionism have been borrowed by modern deep learning techniques. The first concept is represented by the *distributed representation* [19]. According to this principle: (i) each input of a system has to be represented by multiple features; (ii) each of these multiple features needs to be adopted in the representation of many possible inputs. The second concept consists in the adoption of the *back-propagation* algorithm to train deep neural networks [20]. The goal of the back-propagation is to efficiently compute the gradient of the loss function with respect to the weights, or parameters, of a neural network. Through this gradient computation, it is possible to update the network weights to minimize the loss value. As a consequence, the learning process can be reduced to an optimization problem, finding a function that produce the minimal loss. Such methodological advances, in combination with the increasing of computational resources and the availability of large datasets, make the modern deep learning techniques very powerful in being trained without supervision, extracting patterns and regularities from the input data. Making this acquisition process explainable and transparent is still an open research issue in the connectionist community, which is now trying to fill the gap between black-box systems [21] to more understandable systems. Nevertheless, as recognized by some authors (e.g., Lecue [22]), this community is far from creating tools that in terms of explainability adapt to different domains and applications.

Beside the connectionist approach, the movement of symbolic AI, also known as GOFAI (Good Old-Fashioned Artificial Intelligence) [23], adopts an opposed paradigm: the knowledge about the world is not acquired deriving a mathematical model through optimization techniques, but it is hard-coded in the system exploiting formal languages. Therefore, the system is able to reason on the statements expressed in these formal languages through logical inference rules. The modern implementation of symbolic systems is represented by ontologies, formal representation of domain conceptualizations, and Knowledge Graphs (KGs), large networks of entities and relationships relevant to a specific domain, where each node of the graph is an entity and each edge is a semantic relationship connecting two different entities. KGs are explicitly designed to capture the knowledge within domains, integrating and linking data from different phenomena, or different types of representation [24]. The reasoning over these artifacts, as reported by Hoehndorf et al. [25], enables (i) *consistency checking* (i.e., recognizing contradictions between different facts), (ii) *classification* (i.e., defining taxonomies),

(iii) *deductive inference* (i.e., revealing implicit knowledge given a set of facts). Considering their features, symbolic methods are not robust to noise and can not be applied to non-symbolic context where the data is ambiguous. Nevertheless, they offer a data-efficient process by which models can be trained to reason on symbolic contexts and are able to provide background knowledge for deep learning models [25]. Lecue [22] stresses the opportunity of the usage of KGs to encode the semantic of input and output data, considering also their specific properties. Moreover, KGs could play a central role in designing novel deep learning architectures that incorporate causation [12], supporting also other tasks, from data integration and discovery to semantic fragmentation and composition [22]. All these features represent an essential condition to create, as defined by Doran [4], a system that is able to emit comprehensible symbols apart from its output. This condition enables the opportunity for the user to interact with the system, exploiting her tacit form of reasoning and knowledge [4] that relies on shared symbols.

3. Explanations for AI Experts: Technical Issues and Solutions

The goal of an explainable system is to expose intelligible explanations in a human-comprehensible way, and keeping the human in the loop is a determinant aspect. Nevertheless, as clearly demonstrated in the literature survey conducted by Adadi et al. [26], the human factor impact is not adequately considered XAI. Concordantly, Miller et al. [27] argue that most of the existing methods focus on the perspective of technicians rather than the viewpoint of the intended users.

This condition creates a gap of expectations of the layman in terms of explainable systems, because the term "explanation" has been re-purposed by the connectionist community to address specific technical issues. Instead, other disciplines collectively considered as "explanation sciences" [28], from cognitive science to philosophy, are rarely included in the current XAI research.

In the following subsections we discuss the technical issues related to eXplainable AI, and we provide a brief overview of the typical approaches adopted by AI experts for the explanation task.

3.1. Technical Issues in a Connectionist Perspective

We synthesize the technical issue related to the explainability as follows: (i) complexity; (ii) multiplicity; (iii) opacity degree.

Complexity: deep learning techniques are difficult to examine, because of their structure and the way they are working. As reported by Adadi [26], since deep learning algorithms are based on high-degree interactions between input features, the disaggregation of such functions in a human understandable form and with human meaning is inevitably more difficult.

Multiplicity: considering the complex network architecture of deep learning techniques, these techniques might produce multiple accurate models from the same training data. Hall and Gill [29] define this issue as "multiplicity of good models". In fact, the internal paths through the network are very similar, but not the same, and consequently the related explanations can change across different models. This issue clearly emerges in case of *knowledge extraction* techniques [30,31]. The explanation in this case consists in the knowledge acquired and encoded as the internal representation of the network during the training.

Opacity Degree: a well-known problem in the AI research field is the trade off between interpretability and accuracy. As underlined by Breiman [32], achieving high accuracy often requires more complex prediction methods, and vice versa simple and interpretable functions do not provide accurate predictors. Considering this issue, the traditional approach is to construct complex models to reach a high accuracy and then adopt reverse engineering techniques to extract explanations, thus without the necessity of knowing in details the inner works of the original model [26].

3.2. Explainable Systems for AI Experts

Considering the technical issues mentioned in the previous subsection, we identify two main explanation methods from literature, that is, transparency and post-hoc [28], that can support explanations to the AI expert in sensitive conditions.

Transparency Explanations: these types of methods are focused on how a model work internally and, as underlined by Lepri et al. [33], the explanation can be rendered at three different levels: (i) the entire model, (ii) the individual components (e.g., parameters) of the model, (iii) the specific training algorithm. For each level, Mittelstadt et al. [28] recognize respectively the following notions in terms of explanation:

1. **Simulatability**: checking through a heuristic approach whether a human reaches the mechanistic understanding of how the model functions, and consequently if he is able to simulate the decision process. In this context, within a user study [34] that involved thousand participants, Friedler et al. measured human performance in operations that mimic the definition of simulatability, using as evaluation metric the runtime operation count.
2. **Decomposability**: in this case each component of the model, including a single input, parameter, and computation has to be clearly interpretable. In a recent work, Assaf et al. [35] introduce a Convolutional Neural Network (CNN) to predict multivariate time series, in the domain of renewable energy. The goal is to produce saliency maps [36] to provide two different types of explanation on the predictions: (i) which features are the most important in a specific interval of time; (ii) in which time intervals the joint contribution of the features has the greatest impact.
3. **Algorithmic transparency**: for techniques such as linear models there is a margin of confidence that the training will converge to a unique solution, so the model might behave in an online setting in an expected way. At the opposite, deep learning models cannot provide guarantees that they will work in the same way on new problems. Datta et al. [37] designed a set of Quantitative Input Influence (QII) for capturing the joint influence of the inputs on the outputs of an AI system, with the goal to produce transparency reports.

Post-hoc Explanations: these methods do not seek to reveal how a model works, but they are focused on how it behaved and why. Lipton [38] detects different post-hoc approaches that include natural language explanations, interactive visualizations, local explanations, and case-based explanations. Natural language explanations are based on qualitative artifacts that describe the relationships between the features of the input data and the outputs (e.g., predictions or classifications) of the model [39]. Interactive visualizations show relative influence of features or provide graphical user interfaces to explore visual explanations [40]. Local explanations intend to identify the behavior of a deep learning model on a particular prediction in two different ways—a simple and local fitting around a particular decision [39], and variables perturbations to understand the changes in the prediction [41]. Case-based explanations consist in the exploitation of the trained model to identify which samples of the training data are the most similar to the prediction or the decision to be explained [42].

Despite this variety of approaches to address technical issues related to the explainability, these methodologies are not intended to capture the full behavior of the system, but rather to provide approximations of the system behavior. As stressed by Mittelstadt et al. [28], these approximations can be offered to AI experts both for "pedagogical purposes" and to provide reliable predictions on the system behavior over a restricted domain. Nevertheless, technical approximations are not enough and can be misleading when presented as an explanation to the non-insiders on how the model works.

4. Explanations for Non-Insiders: Three Research Challenges with Symbolic Systems

In his prominent work, Miller [27] defines explanations as social conversation and interaction for transfer knowledge. A fruitful exchange implies that who explains must be able to recognize the mental model of who receives the information. To enable this process, a representation of the world

through symbols is a crucial requirement. Furthermore, Miller articulates that this social exchange can be enabled in the XAI only if human sciences, such as philosophy, psychology, and cognitive sciences are injected within the development of new XAI approaches. The final aim is in fact is to produce explanations that allow "affected parties, regulators and more broadly non-insiders to understand, discuss, and potentially contest decisions provided by black-box models" [26]. Considering these requirements, we believe that symbolic systems open opportunities in three different human-centric challenges: (i) knowledge matching, (ii) cross-disciplinary explanations and (iii) interactive explanations.

The identification of these three challenges is based on our analysis, synthesis and elaboration of the most recent advances in the field of incorporating KGs features into deep learning model. We revised such corpus of evidence in accordance with the aim of building AI systems able to provide comprehensible explanations for non-insiders, through manipulation of symbols. Herein we select, for each challenge, recent research works that in our opinion represent the most promising references. Therefore, we suggest that further work along these tracks should be encouraged and supported.

Knowledge Matching: Seeliger et al. [10] identify the matching of input features or internal neurons of deep learning models to classes of an ontology or entities of a KG as an important challenge to be addressed by the research community of XAI. Interesting works in this sense have been proposed by Sarker et al. [43], in which objects within images are mapped to the classes of the Suggested Upper Match Ontology. On the basis of the classification output of the neural network, a description logic (DL) learner is adopted to create class expressions that work as explanations. In a recent work, Angelov et al. [44] introduce a novel deep learning architecture for image classification, that includes an hidden semantic layer, called "Prototype layer", to provide a clear explanation of the model. This intermediate semantic layer receives training data samples, defined as *prototypes*, that are characterized by local peaks in the data distribution. Each prototype is then exploited to generate logical rules to provide natural language explanations. From a different perspective, Selvaraju et al. propose a method to learn a map between neurons weight to semantic domain knowledge [45]. In a work more focused on unsupervised learning, Batet et al. [46] exploit WordNet [47] and its taxonomic knowledge to compute semantic similarities that conduct to more interpretable clusters. In the context of transfer learning, Geng et al. [48] exploit two external KGs in a deep learning architecture for the following purposes: (i) provide explanations to understand the transferability of features learned by a CNN between different domains; (ii) justify new classes predicted by a Graph Convolutional Network (GCN), that were unseen by the CNN.

Cross-disciplinary Explanations: ontologies and KGs are able to represent domains by means of symbols, whose manipulation produces transparent inferences of new information. Both implementations are able to outline different areas of human knowledge according to characteristics and expertise of the related users. Moreover, it is worth mentioning that the concept of ontology is adopted by information science and philosophy. The stretch in common within the two different disciplines is the attempt to define ideas and entities within a precise system of categories, that explicit interdependent properties and relationships. Therefore, applied ontologies can be considered a technical application to prior work in philosophy. In a work entitled "The Knowledge Graph as the Default Data Model for Machine Learning", Wilcke et al. [49] describe how decades of work have been devoted to the development of vast and distributed KGs to represent various domains of knowledge. Potentially, the usage of this form of interlinked and structured data enables the training of deep learning models from different domains and from the perspective of different disciplines. In this context, Sopchoke et al. [50] developed a method for explainable rules in recommendation systems related to different domains, using relational learning techniques on semantic statements.

Interactive Explanations: in a human-centered vision of explainable tools, AI systems should be able to offer user interaction features rather than static explanations. Wang et al. [51] developed a neural network that extracts image contents as KG facts that are interlinked with the DBpedia repository [52], and questions provided by the user are translated in SPARQL (SPARQL Protocol and

RDF Query Language) queries that are run over this enriched knowledge base. Liao et al. [53] propose a recommendation system that enable user-feedback on human-interpretable domain concepts. More in general, in recommendation systems the ontology is able to provide information that is implicit in the data used to perform inference and consequently to create rules which limit the number of plausible recommendations. For future developments, Sarker et al. [43] envision their explanation tool for image classification to be used in an interactive system where a human can monitor and fix algorithmic decisions based on the given explanations.

5. Conclusions and Future Work

Explainability of the outputs of AI systems is an essential requirement in domains where their impact on human life is relevant. Nevertheless, the leading implementation of modern AI based on deep learning model is barely intelligible to the layman, as well as the main technical solutions proposed in the field of explainable AI are usable only by experts of the field. A promising alternative to most beaten routes for explainable AI is represented by Knowledge Graphs, because they are natively developed to support explanations intelligible to humans. In this work we analyzed the differences in terms of learning approaches between deep learning techniques, belonging to the connectionist movement, and KGs and ontologies, expression of current symbolic systems. On the basis of this analysis, we provided an overview of the main issues and approaches to address XAI in the connectionist communities, which are primarily aimed at AI experts. We state that for creating a human-centered AI able to emit and manipulate symbols that are comprehensible also for non-insiders, symbolic systems has to be integrated in the dominant approaches of AI, because they can contribute in terms of (i) *knowledge matching*, (ii) *cross-disciplinary explanations* and (iii) *interactive explanations*. In particular, the knowledge matching of deep learning components, including input features, hidden unit and layers, and output predictions with KGs and ontology components can make the internal functioning of algorithms more understandable. Moreover, query and reasoning mechanisms of KGs and ontologies enable the conditions for cross-disciplinary and interactive explanations. For each of the three tracks, we provided references to the most recent and prominent -in our opinion- research works. We suggest that further work along these tracks should be encouraged and supported to make explanations of AI systems outputs more inclusive and effective. Starting from these points we identify specific trajectories for future work on XAI, including the exploitation of symbolic techniques to design novel deep neural architectures to natively encode explanations; the development of multi-modal explanation models that are able to provide insights from different perspectives, combining visual and textual artifacts; the definition of a common explanation framework for the deep learning model comparison, based on KGs and ontologies, to enable proper validation strategies.

Author Contributions: Conceptualization, G.F. and A.V.; Investigation, G.F. and A.V.; Methodology, G.F. and A.V.; Validation, G.F. and A.V.; Writing—original draft, G.F. and A.V.; Writing—review & editing, G.F. and A.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors would like to thank Marco Conoscenti, post-doc researcher at the Nexa Center for Internet and Society, and Giovanni Garifo, research fellow at the Nexa Center for Internet and Society, for constructive criticism of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lv, Y.; Duan, Y.; Kang, W.; Li, Z.; Wang, F.Y. Traffic flow prediction with big data: A deep learning approach. *IEEE Trans. Intell. Transp. Syst.* **2014**, *16*, 865–873. [CrossRef]
2. Ding, X.; Zhang, Y.; Liu, T.; Duan, J. Deep learning for event-driven stock prediction. In Proceedings of the Twenty-Fourth International Joint Conference on Artificial intelligence, Buenos Aires, Argentina, 25–31 July 2015.

3. Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity prediction using deep learning. *Front. Environ. Sci.* **2016**, *3*, 80. [CrossRef]
4. Doran, D.; Schulz, S.; Besold, T.R. What does explainable AI really mean? A new conceptualization of perspectives. *arXiv* **2017**, arXiv:1710.00794.
5. Saradhi, V.V.; Palshikar, G.K. Employee churn prediction. *Expert Syst. Appl.* **2011**, *38*, 1999–2006. [CrossRef]
6. Kang, H.W.; Kang, H.B. Prediction of crime occurrence from multi-modal data using deep learning. *PLoS ONE* **2017**, *12*, e0176244. [CrossRef]
7. Al Hasan, M.; Chaoji, V.; Salem, S.; Zaki, M. Link prediction using supervised learning. In Proceedings of the SDM04: Workshop on Link Analysis, Counter-Terrorism and Security, Lake Buena Vista, FL, USA, 24 April 2004.
8. Lee, H.; Yune, S.; Mansouri, M.; Kim, M.; Tajmir, S.H.; Guerrier, C.E.; Ebert, S.A.; Pomerantz, S.R.; Romero, J.M.; Kamalian, S.; et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nat. Biomed. Eng.* **2019**, *3*, 173. [CrossRef]
9. Ehrlinger, L.; Wöß, W. Towards a Definition of Knowledge Graphs. SEMANTiCS (Posters, Demos, SuCCESS) 2016. Available online: https://2016.semantics.cc/posters-and-demos-madness (accessed on 21 February 2020).
10. Seeliger, A.; Pfaff, M.; Krcmar, H. Semantic Web Technologies for Explainable Machine Learning Models: A Literature Review. PROFILES 2019. Available online: http://ceur-ws.org/Vol-2465/ (accessed on 21 February 2020).
11. Chekol, M.W.; Pirrò, G.; Schoenfisch, J.; Stuckenschmidt, H. Marrying uncertainty and time in knowledge graphs. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
12. Munch, M.; Dibie, J.; Wuillemin, P.H.; Manfredotti, C. Interactive Causal Discovery in Knowledge Graphs. 2019. Available online: http://ceur-ws.org/Vol-2465/ (accessed on 21 February 2020).
13. Gruber, T.R. Toward principles for the design of ontologies used for knowledge sharing? *Int. J. Hum.-Comput. Stud.* **1995**, *43*, 907–928. [CrossRef]
14. Dumontier, M.; Callahan, A.; Cruz-Toledo, J.; Ansell, P.; Emonet, V.; Belleau, F.; Droit, A. Bio2RDF release 3: A larger connected network of linked data for the life sciences. In Proceedings of the 2014 International Conference on Posters & Demonstrations Track, Riva del Garda, Italy, 21 October 2014; Volume 1272, pp. 401–404.
15. Banerjee, S. A Semantic Web Based Ontology in the Financial Domain. In *Proceedings of World Academy of Science, Engineering and Technology*; World Academy of Science, Engineering and Technology (WASET): Paris, France, 2013; Number 78, p. 1663.
16. Casanovas, P.; Palmirani, M.; Peroni, S.; van Engers, T.; Vitali, F. Semantic web for the legal domain: The next step. *Semant. Web* **2016**, *7*, 213–227. [CrossRef]
17. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
18. Rumelhart, D.E. Parallel distributed processing: Explorations in the microstructure of cognition. *Learn. Intern. Represent. Error Propag.* **1986**, *1*, 318–362.
19. Hinton, G.E.; McClelland, J.L.; Rumelhart, D.E. *Distributed Representations*; Carnegie-Mellon University: Pittsburgh, PA, USA, 1984.
20. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
21. Koh, P.W.; Liang, P. Understanding black-box predictions via influence functions. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 7–9 August 2017; Volume 70, pp. 1885–1894.
22. Lecue, F. On the Role of Knowledge Graphs in Explainable AI. Semantic Web Journal (Forthcoming). 2019. Available online: http://www.semantic-web-journal.net/content/role-knowledge-graphs-explainable-ai (accessed on 21 February 2020).
23. Haugeland, J. *Artificial Intelligence: The Very Idea*; MIT Press: Cambridge, MA, USA, 1989.
24. Heath, T.; Bizer, C. Linked data: Evolving the web into a global data space. *Synth. Lect. Semant. Web Theory Technol.* **2011**, *1*, 1–136. [CrossRef]
25. Hoehndorf, R.; Queralt-Rosinach, N. Data science and symbolic AI: Synergies, challenges and opportunities. *Data Sci.* **2017**, *1*, 27–38. [CrossRef]

26. Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
27. Miller, T.; Howe, P.; Sonenberg, L. Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *arXiv* **2017**, arXiv:1712.00547.
28. Mittelstadt, B.; Russell, C.; Wachter, S. Explaining explanations in AI. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; ACM: New York, NY, USA, 2019; pp. 279–288.
29. Hall, P.; Gill, N. *An Introduction to Machine Learning Interpretability-Dataiku Version*; O'Reilly Media, Incorporated: Sebastopol, CA, USA, 2018.
30. Johansson, U.; König, R.; Niklasson, L. The Truth is In There-Rule Extraction from Opaque Models Using Genetic Programming. In Proceedings of the International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004, Miami Beach, FL, USA, 17–19 May 2004; pp. 658–663.
31. Sadowski, P.; Collado, J.; Whiteson, D.; Baldi, P. Deep learning, dark knowledge, and dark matter. In Proceedings of the NIPS 2014 Workshop on High-energy Physics and Machine Learning, Montreal, QC, Canada, 13 December 2015; pp. 81–87.
32. Breiman, L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Stat. Sci.* **2001**, *16*, 199–231. [CrossRef]
33. Lepri, B.; Oliver, N.; Letouzé, E.; Pentland, A.; Vinck, P. Fair, transparent, and accountable algorithmic decision-making processes. *Philos. Technol.* **2018**, *31*, 611–627. [CrossRef]
34. Friedler, S.A.; Roy, C.D.; Scheidegger, C.; Slack, D. Assessing the local interpretability of machine learning models. *arXiv* **2019**, arXiv:1902.03501.
35. Assaf, R.; Schumann, A. Explainable deep neural networks for multivariate time series predictions. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019; pp. 6488–6490.
36. Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; Kim, B. Sanity checks for saliency maps. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 9505–9515.
37. Datta, A.; Sen, S.; Zick, Y. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In Proceedings of the 2016 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2016; pp. 598–617.
38. Lipton, Z.C. The mythos of model interpretability. *arXiv* **2016**, arXiv:1606.03490.
39. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA, 2016; pp. 1135–1144.
40. Tamagnini, P.; Krause, J.; Dasgupta, A.; Bertini, E. Interpreting black-box classifiers using instance-level visual explanations. In Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics, Chicago, IL, USA, 14 May 2017; ACM: New York, NY, USA, 2017; p. 6.
41. Adler, P.; Falk, C.; Friedler, S.A.; Rybeck, G.; Scheidegger, C.; Smith, B.; Venkatasubramanian, S. Auditing black-box models by obscuring features. *arXiv* **2016**, arXiv:1602.07043.
42. Kim, B.; Rudin, C.; Shah, J.A. The bayesian case model: A generative approach for case-based reasoning and prototype classification. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 1952–1960.
43. Sarker, M.K.; Xie, N.; Doran, D.; Raymer, M.; Hitzler, P. Explaining trained neural networks with semantic web technologies: First steps. *arXiv* **2017**, arXiv:1710.04324.
44. Angelov, P.; Soares, E. Towards Explainable Deep Neural Networks (xDNN). *arXiv* **2019**, arXiv:1912.02523.
45. Selvaraju, R.R.; Chattopadhyay, P.; Elhoseiny, M.; Sharma, T.; Batra, D.; Parikh, D.; Lee, S. Choose Your Neuron: Incorporating Domain Knowledge through Neuron-Importance. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 526–541.
46. Batet, M.; Valls, A.; Gibert, K.; Sánchez, D. Semantic Clustering Using Multiple Ontologies. In Proceedings of the Catalan Conference on AI (CCIA), Terragona, Spain, 20–22 October 2010, pp. 207–216.
47. Miller, G.A. WordNet: A lexical database for English. *Commun. ACM* **1995**, *38*, 39–41. [CrossRef]
48. Geng, Y.; Chen, J.; Jimenez-Ruiz, E.; Chen, H. Human-centric transfer learning explanation via knowledge graph. *arXiv* **2019**, arXiv:1901.08547.

49. Wilcke, X.; Bloem, P.; de Boer, V. *The Knowledge Graph as the Default Data Model for Machine Learning*; IOS Press: Amsterdam, The Netherlands, 2017.
50. Sopchoke, S.; Fukui, K.i.; Numao, M. Explainable cross-domain recommendations through relational learning. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
51. Wang, P.; Wu, Q.; Shen, C.; Hengel, A.V.; Dick, A. Explicit knowledge-based reasoning for visual question answering. *arXiv* **2015**, arXiv:1511.02570.
52. Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P.N.; Hellmann, S.; Morsey, M.; Van Kleef, P.; Auer, S.; et al. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semant. Web* **2015**, *6*, 167–195. [CrossRef]
53. Liao, L.; He, X.; Zhao, B.; Ngo, C.W.; Chua, T.S. Interpretable multimodal retrieval for fashion products. In Proceedings of the 2018 ACM Multimedia Conference on Multimedia Conference, Seoul, Korea, 22–26 October 2018; pp. 1571–1579.