Semantic Web 1 (0) 1–5
IOS Press

On The Role of Knowledge Graphs in Explainable AI

Freddy Lecue [a,b],
[a] *CortAIx, Thales, Montreal, Canada*
E-mail: freddy.lecue.e@inria.fr
[b] *WIMMICS, Inria, Sophia Antipolis, France*
E-mail: freddy.lecue.e@inria.fr

Editors: First Editor, University or Company name, Country; Second Editor, University or Company name, Country
Solicited reviews: First Solicited Reviewer, University or Company name, Country; Second Solicited Reviewer, University or Company name, Country
Open reviews: First Open Reviewer, University or Company name, Country; Second Open Reviewer, University or Company name, Country

Abstract. The current hype of Artificial Intelligence (AI) mostly refers to the success of machine learning and its sub-domain of deep learning. However AI is also about other areas such as knowledge representation and reasoning, or distributed AI i.e., areas that need to be combined to reach the level of intelligence initially envisioned in the 1950s. Explainable AI (XAI) is now referring to the core backup for industry to apply AI in products at scale, particularly for industries operating with critical systems. This paper reviews XAI not only from a Machine Learning perspective, but also from the other AI research areas such as AI Planning or Constraint Satisfaction and Search. We expose the XAI challenges of AI fields, their existing approaches, limitations and opportunities for knowledge graphs and their underlying technologies.

Keywords: Knowledge graph, explainable AI

1. Introduction

Artificial Intelligence (AI), as a discipline aiming at building intelligent machines mimicking "cognitive" functions that humans associate with other human minds, such as "learning" and "problem solving" [1], is addressing intelligence for systems from a large variety of facets. From machine learning (ML) to knowledge representation and reasoning (KRR), game theory, uncertainty in AI (UAI), robotics, multi-agent systems, constraint satisfaction and search (CSS), planning and scheduling, computer vision, natural language processing, all are foundational pillars of the AI as we know it today. All latter sub-fields of AI have matured, specialized, and sometimes converged together with the aim of accessing to general artificial intelligence i.e., the holy grail of AI.

Many research questions have been vertical to all sub-fields of AI such as decidability and complexity from a theoretical perspective or scalability from a more applicability dimension. However one is remaining current, even getting more traction than others in the new world of industrialized AI: explainability. Obtaining explainable AI systems consists in addressing the following question: “how to build intelligent systems able to expose explanation in a human-comprehensible way” for any of its AI decision. We will use the well-adopted XAI term, standing for eXplainable AI, when referencing to the explanation problem in AI. Answering this XAI question is far from trivial, and has been studied for years in all subfields of AI, with no exception. Such problem has been tackled under different names, concepts, definitions, with various requirements and objectives. For instance interpretation and justification are terms coined in KRR, diagnostics in UAI, debugging in robotics, constraints relaxation in CSS, features importance in ML, or features attribution for Neural Networks [2, 3].

Despite a surge of innovation focusing on ML-based AI systems such question of explainability has not been deeply studied as much as in the other AI subfields such as KRR. However answers to this question of explainability and questions related to the responsibility, validity (e.g., robustness), privacy-preserving and more broadly trust of AI systems (Figure 1) will be intrinsically connected to the adoption of AI in industry at scale, particularly in industries operating with critical systems. Indeed explanation, which could be used for debugging intelligent systems or deciding to follow a recommendation in real-time, will increase acceptance and user trust.

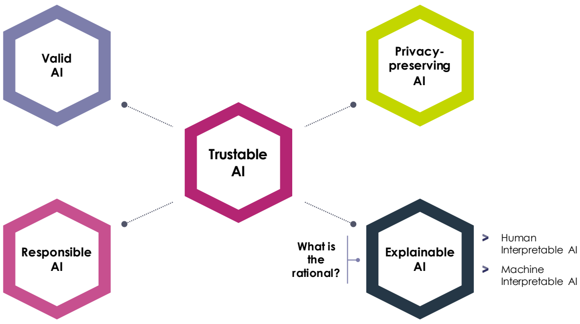

Fig. 1. On the Combination of Valid, Responsible, Privacy-preserving and Explainable AI towards Trustable AI.

Unsurprisingly, the exact same research community, from which emerged the most successful ML-based AI systems [4, 5], is now trying to fill the gap between black-box ML systems [6] to more white-box ML systems. Some approaches are most successful than others, but still the AI community is far from having self-explainable AI systems which automatically adapt to any (i) data, (ii) ML algorithm, (iii) model, (iv) user, or (v) application and (v) context. Even more surprisingly, only a few work in KRR and its subfields of Web and AI i.e., semantic Web [7], linked data [8], and more recently knowledge graphs [9], engaged in the endeavour of explaining the broader family of ML-based systems. However KRR, the semantic Web together with knowledge graphs, aiming at representing and reasoning over structured information, should be designed and armed to move XAI closer to human comprehension.

This paper reviews XAI in the various fields of AI i.e., by first describing the main research question, its XAI challenge, existing approaches, their limitations and opportunities for knowledge graphs and their underlying technologies.

2. Knowledge Graph for XAI Methods

This section highlights the main research question in major AI fields, their associated XAI challenge (Figure 2), together with existing approaches, their limitations and opportunities for semantic Web and knowledge graph technologies. AI areas are broken down following the AAAI taxonomy for research paper submission [10]. Although such a taxonomy has some limitations e.g., arbitral limits, natural intersection of AI domains, at least it benefits from a well-accepted list of fields in AI, which are well-represented in major generalist AI conferences such as IJCAI [11] and ECAI [12].

2.1. Machine Learning (except Neural Netwok)

• ***Research Question***: ML algorithms [13] aim at elaborating a mathematical model based on sample data, known as“training data", in order to make predictions or decisions on unseen data, known as “test data" without being explicitly programmed to perform the task. Three main tasks of learning are studied: (i) supervised learning if data contains both input and labeled data, (ii) unsupervised learning to derive some structures in data if labels are not exposed, and (iii) reinforcement learning if further information could be captured through interaction with the environment.

• ***XAI Challenge***: All tasks of ML expose a mathematical models through an appropriate, but somehow abstract representation of data. XAI in ML [14] is about explanation of (i) models, known as global explanation, and (ii) a prediction, known as local explanation.

• ***Approaches***: Some models are naturally designed to explicit their rational e.g., linear regression, decision trees, generalized linear (or additive), naive bayes models. In case of more complex models, some of their representative elements such as feature importance, partial dependency plot or individual conditional expectation can be used for capturing high level representation of the ML model for global explanation. State-of-the-art approaches [15, 16] go further by revisiting feature importance for local explanation.

• ***Limitations***: Most approaches limits explanation to features involved in the data and model, or at best to examples, prototypes [17] or counterfactuals [18]. Explanation should go beyond correlation (which is what features importance is about) and numerical similarity (which is what local explanation is about).

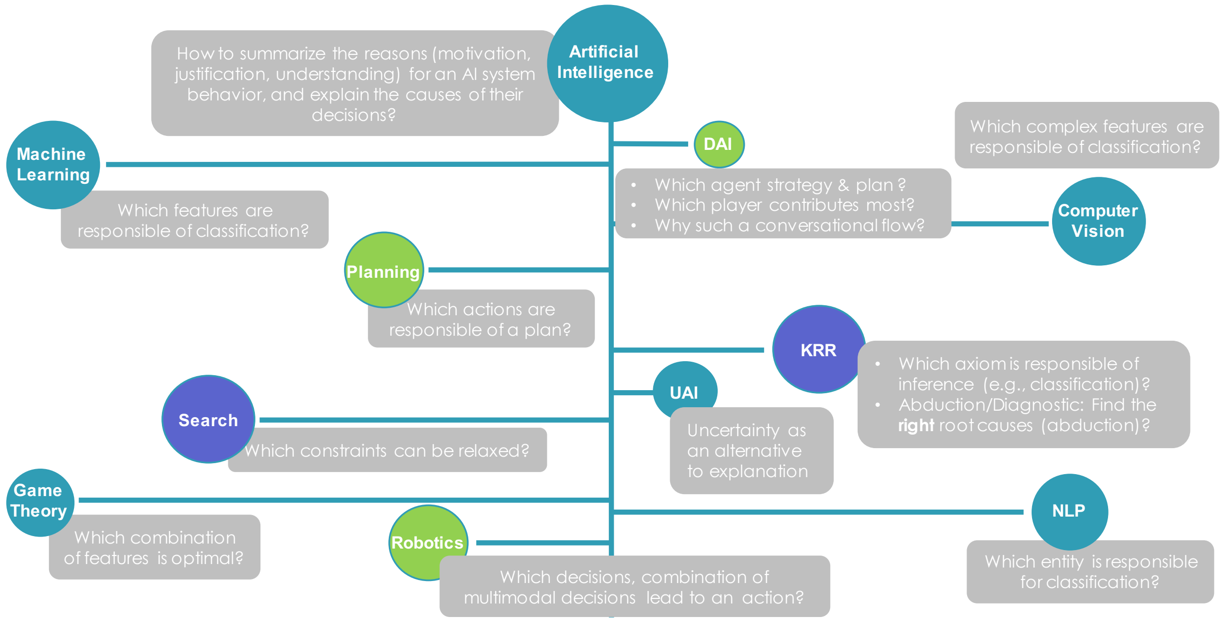

Fig. 2. XAI Challenges in Major AI Fields. (DAI: Distributed AI, UAI: Uncertainty in AI, KRR: Knowledge Representation and Reasoning, NLP: Natural Language Processing)

• ***Opportunity***: Knowledge graphs do encode contexts, expose connections and relations, and support inference and causation natively. Existing XAI approaches in ML consider a flat representation of data, and context is out of the loop of the explanation process. Knowledge graphs could be used for encoding better representation of data, structuring a ML model in a more interpretable way, adopt semantic similarity for local explanation. In addition we could envision approaches relying on knowledge graphs to compact large trees in decisions trees or forrest. For instance combinations of nodes could be captured as a unique (probabilistic) concept or property in a knowledge graph.

2.2. Artificial (Deep) Neural Network

• ***Research Question***: Similarly to other ML approaches, Artificial Neural Network (ANN) aims at learning representation. The main differentiator with other approaches is its scalability and performance with high number of features and instances, which fit better images and texts.

• ***XAI Challenge***: Both local and global explanations are strong focus of the ANN community.

• ***Approaches***: Contrary to other ML approaches, there is no easy way around explanation of ANN models or predictions. Existing techniques either encode feature importance through attribution [2, 3], attention mechanism [19], or obtain a more interpretable approximation through surrogate models [20] such as decision tree.

• ***Limitations***: Explanations are artificially built, for instance by forcing the network to focus on some group of features or correlations at best. In addition they do not represent any logic of the learning task, making explanation a very difficult task to achieve. The latter is due to the foundational theory of ANN, which consists in deriving a mathematical model through local optimizations.

• ***Opportunity***: Novel ANN architectures needs to be designed to natively encode explanation. Some recent approaches which aims at capturing better model hierarchical relationships [21], or causality mechanism [22] are promising. However they could be polished further by (i) adding logic representation layers in ANN, such as [23] using network dissection approaches [24], (ii) encoding the semantics of inputs, outputs and their properties cf. Figure **??**. Knowledge graphs could play a central roles in such a new design, particularly as novel architectures should embed causation and feature reasoning. Such design could advance ANN further by supporting integration, discovery, fragmentation, or composition.

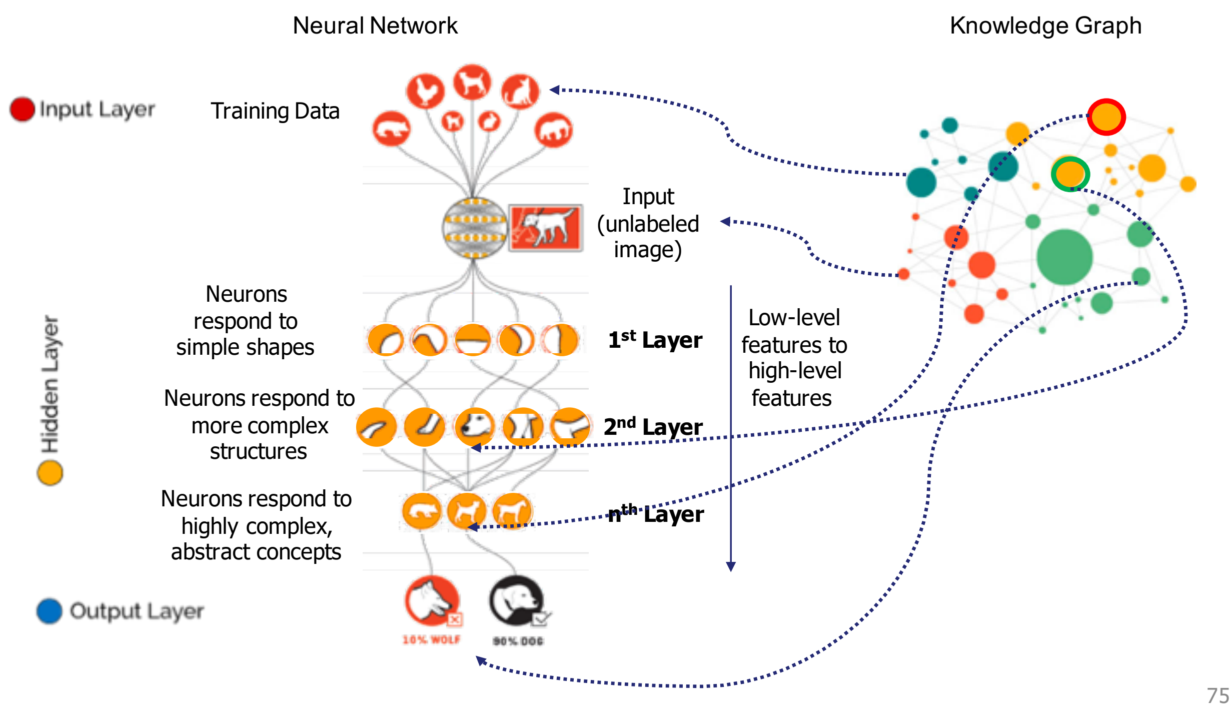

Fig. 3. On the Role of Knowledge Graphs for Explainable Artificial (Deep) Neural Network. (What is the causal relationship between the input / output / training data?)

2.3. Computer Vision

• ***Research Question***: Computer vision is relying on ANN architectures due to the nature and size of its data. Tasks range from semantic segmentation, object detection, scene reconstruction, visual question answering.

• ***XAI Challenge***: The main XAI task in computer vision is identification of pixels, or group of pixels responsible for triggering a shape detection, an uncertainty or an error. Explanation is often referred as visual inspection due to the nature of data processed.

• ***Approaches***: Saliency maps [25] are classic methodologies in computer vision. They include many variant of gradient modification for capturing representative features. Network dissection [24] is another approach segmenting ANN to derive interpretable units and layers.

• ***Limitations***: Although saliency map expose interesting visualization artifacts, they do not capture any semantics. At best those artifacts capture a disentangled representation, which remains subject to human interpretation. Knowledge graphs could expose the semantics of such disentangled representation. However integrating semantics in ANN, hidden units of feature space remain open challenges.

• ***Opportunity***: Adding semantics could help answering other open questions[1] such as: What is a disentangled representation, and how can its factors be quantified and detected? Do interpretable hidden units reflect a special alignment of feature space, or are interpretations a chimera? What conditions in state-of-the-art training lead to representations with greater or lesser entanglement? What is the semantics of a group of hidden units in a neural network?

2.4. Constraint Satisfaction and Search

• ***Research Question***: Constraint satisfaction and Search aims at finding a solution to a set of constraints that impose conditions that the variables must satisfy. A solution is a set of values for the variables that satisfies all constraints. Constraints are defined on a finite domain.

• ***XAI Challenge***: The main challenge is to identify which constraints to relax for conflict resolutions. Ex-

[1] http://netdissect.csail.mit.edu/

planations are usually a subset of variables which satisfies a set of constraints.

• ***Approaches***: Constraint satisfaction problems on finite domains are typically solved using a form of search. Backtracking, constraint propagation, local search are examples of such approaches. Even though the problem is known to be a NP complete problem with respect to the domain size, research has shown a number of tractable sub-cases with promising approaches [26], [27].

• ***Limitations***: Even though optimal structures and search spaces have been largely introduced in the community, complexity remains one of the main limitations.

• ***Opportunity***: It has been demonstrated that any structure in problem representation has largely benefited search [28]. We could envision more knowledge-driven structure, inspired from knowledge graphs, which could dynamically adapt to variables, constraints, search space. Knowledge graphs could even drive search through semantic and logical relations among constraints, which could be modelled as entities in a graph.

2.5. Game Theory

• ***Research Question***: Game theory [29] is the study of mathematical models of strategic interaction between rational decision-makers. Examples of games include zero-sum games [30], in which one person's gains result in losses for the other participants.

• ***XAI Challenge***: Game theory has been dealing with XAI from its inception as one of its main challenge is to identify and to understand the underlying mathematical model as well as its properties. Game theory is applied to a wide range of behavioural relations, and is now an umbrella term for the science of logical decision making in humans, animals, and computers, in which explanation is the core question driving the modelling.

• ***Approaches***: The Shapley value [31] is a solution concept in game theory, which inspired recent research in Machine Learning to address the problem of explanation [16]. The Shapley value is characterized by a collection of desirable properties, and is used to capture the influence of a player in a game settings (or a feature in a machine learning setting). Such properties characterize the explanation.

• ***Limitations***: Similarly to the domain of constraint satisfaction and search, complexity is a challenge for explainability in game theory. Only an approximate solution is feasible, usually identified through some randomization feature values coalition.

• ***Opportunity***: As recently explored structured representation of the models as its features [32] has shown better scalability, while not necessarily improving explainability. Knowledge graphs could be considered to better structure models, organize features, then reducing the search space and potentially improve understanding and readability of explanation, particularly when embedded in a structured set of connected entities.

2.6. Uncertainty in AI

• ***Research Question***: The field of Uncertainty in AI is at the frontier of various AI fields, namely knowledge representation, learning and reasoning. Bayesian probability is one of the core fundamental, and Probabilistic Graphical Models (PGMs) [33] are usually central for representing and reasoning with uncertainty as they encode probability distributions.

• ***XAI Challenge***: Graphical models are often used to model multivariate data, since they allow to represent high-dimensional distributions compactly. The explanations draw their attention on the compact distributions and their underlying data. Explanation is then naturally embedded through those relationships, usually through interdependencies and decomposition in data.

• ***Approaches***:

[34] Some approaches are formulating PGMs as weighted logical formulas [35] to tightly decouple the constraints and dependencies from the probabilistic parameters. Reasoning can then be performed on the logic representations. Other approaches analyzes latent spaces and its direct connections with the underlying data [36]. The strength of existing approaches is the underlying reasoning capabilities that PGMs and other probabilistic and logic systems offer.

• ***Limitations***: Even though PGMs are appropriate representations to connect inter-dependable data, dependencies remains probabilistic. Therefore humans are required to remain in the loop to interpret any dependencies. Even embedded in logical formulas there is little gained as we are still embedded in the framework of standard probability theory.

• ***Opportunity***: Semantic representations and connections through knowledge graphs could be used to disambiguate and force latent variables to represent interpretable content.

2.7. Robotics

• ***Research Question***: Robotics is an interdisciplinary branch of engineering and AI science, which deals with the design, construction, operation, and use of robots, as well as computer systems for their control, sensory feedback, and information processing. The underlying technologies are used to develop machines that can replicate human actions. They usually combine and integrate many of the technologies in the AI field.

• ***XAI Challenge***: XAI is required in Robotics mainly for debugging and resolving discrepancy between a solution and an expected answer. Some of the XAI challenges are (1) the rational of coordination in multi-robots Systems and swarms, (2) the fusion of explanation coming from many underlying AI systems such as planning, computer vision, or reasoning. They are unique challenges for robotics with many interesting opportunities as explanation is multi-modal, could be complementary but also conflicting, is spatial and temporal, is driven by goals but also initial conditions.

• ***Approaches***: Narration of autonomous robot experience [37] together with approaches of summarization [38] have been recently introduced as a succinct way of presenting the decision process of robots. Various levels of granularity in the decision process are provided.

• ***Limitations***: Although the latter models extract information from a large poll of data, such systems do not explain their actions and justify their decisions [39]. Explanation is usually to fine-grained to be properly integrated by humans. Seamless integration of multi-modal explanation is also not addressed in the literature.

• ***Opportunity***: The level of abstraction in explanation together with its multi-modal fusion are net opportunities for knowledge graphs. Some semantics could deeply support in exposing appropriate and personalized representations of explanations while fusing explanation content in a compact and comprehensible representation. Knowledge graphs have been designed to capture knowledge from heterogenous domains, making them a great candidate to achieve explanation per se in robotics.

2.8. Distributed AI

• ***Research Question***: Distributed AI is the field of AI dedicated to the development of distributed solutions for problems. It is related to Multi-Agent Systems but also to any representation, structure, system which could make AI scalable.

• ***XAI Challenge***: Main XAI challenges are focusing on explaining and resolving agent conflicts, based on their intentions and beliefs [40]. State-of-the-art aims at identifying the best strategy, through explanation, to achieve a goal. More recent works focus on human comprehension of agent behaviour, its strategy, and its convergence in case of conflicting intentions and beliefs of agents [41, 42].

• ***Approaches***: Approaches such as [43] determines the motivation for a decision by recalling the situation in which the decision was made, and replaying the decision under variants of the original situation. In such scenario they are able to discover what factors led to the decisions, and what alternatives might have been chosen had the situation been slightly different. Approaches tend to be very close to counterfactual [44] and case-based reasoning [45].

• ***Limitations***: Even though ontology is a core representation layer for agents to communicate and negotiate, it is rarely used for explaining agent behaviour, its strategy and success. Lighter knowledge representations might be envisioned.

• ***Opportunity***: The dynamics of agents interaction should be captured more formally, and embedded with broader common sense knowledge to identify human interpretable explanation. Formalization does not need to be complex. For instance some dedicated knowledge graphs could be used to contextualize the agents environment.

2.9. Automated Planning and Scheduling

• ***Research Question***: Automated planning and scheduling [46] is a branch of artificial intelligence that is about the realization of strategies or action sequences, typically for execution by intelligent agents, autonomous robots and unmanned vehicles. Unlike classical control and classification problems, the solutions are complex and must be discovered and optimized in multidimensional space. It could be done in real-time i.e., on-line, or at design-time i.e., off-line. Solutions usually resort to iterative trial and error processes.

• ***XAI Challenge***: XAI challenges in AI planning [47] are as follows: explaining (i) causal relationships of actions, (ii) why some actions are chosen in particular situations, (iii) why plans are better than some, (iv) why plans could not be computed, (v) why replanning might be required.

• ***Approaches***: Past work on explanations primarily involved the AI system explaining the correctness of its plan and the rationale for its decision in terms of its own model [48].

• ***Limitations***: Existing approaches fail in exposing human-understandable explanation, as rational is usually is limited to the planner's domain e.g., in term of actions and initial situation. This strongly limits the comprehension to experts in the given tasks.

• ***Opportunity***: Knowledge graph could be a way forward to better contextualize complex terms, and even better summarize complex actions in more succinct and meaningful way.

2.10. Natural Language Processing

• ***Research Question***: Natural Language Processing is concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data. Research questions includes (visual [49], multi-turn [50]) question answering [51], conversational agents with broader questions related to speech recognition, natural language understanding and generation.

• ***XAI Challenge***: Similarly to machine learning, identifying importance of feature or entity is critical, as it aims at identifying which part of speech is driving the most relevant information. Other core XAI tasks include: explaining the rational of questions sequencing in dialogue, debugging a plan-based dialogue system [52] or explaining the utterances which were intended to achieve [53]

• ***Approaches***: The problem of identifying the most representative entities in a text classification task is addressed by [15] with many variants. Some works [54] extract plan-based model to understand intention and explain rational of the discourse.

• ***Limitations***: On the one hand ML-based approaches, which focus on important entities in text, suffer from having statistics-based explanation only i.e., mainly based on co-occurrence and correlation. On the other hand plan-based models have not been deeply explored, and many research questions related to their representation, rational in questions sequencing remain open.

• ***Opportunity***: Semantics could support for representation purpose. Knowledge graphs could provide the semantic layer missing from brute-force machine learning approaches on text. They could also drive or at least guide sequencing of questions by refining, abstracting or instantiating obscure terms in questions.

3. Conclusion

Despite a surge of innovation focusing on ML-based AI systems, industry is facing the dilemma of applying in products at scale, particularly for industries operating with critical systems. Trust, and trust in AI has been revelled as the one term coining industry needs to move to the next step. Trustable AI is about responsibility validity, privacy-preserving modelling and also explainability. Explanation, which could be used for debugging intelligent systems or deciding to follow a recommendation in real-time, will increase acceptance and user trust. Explanation in AI has different open questions, meaning, definitions and approaches, depending of which AI fields is touching the question. Although various solutions have been introduced, the question remain open in all areas of AI. We presented their challenges in more details, some of their existing approaches, their limitations and opportunities for knowledge graphs to bring explainable AI to the right level of semantics and interpretability. Indeed significant progress in complex AI tasks such as explainable AI could only be achieved through combinations with semantic layers, empowering explanation of complex AI systems.

References

[1] S.J. Russell and P. Norvig, *Artificial Intelligence - A Modern Approach (3. internat. ed.)*, Pearson Education, 2010. ISBN 978-0-13-207148-2. http://vig.pearsoned.com/store/product/1,1207,store-12521_isbn-0136042597,00.html.

[2] M. Sundararajan, A. Taly and Q. Yan, Axiomatic Attribution for Deep Networks, in: *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, 2017, pp. 3319–3328. http://proceedings.mlr.press/v70/sundararajan17a.html.

[3] A. Shrikumar, P. Greenside and A. Kundaje, Learning Important Features Through Propagating Activation Differences, in: *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August*

2017, 2017, pp. 3145–3153. http://proceedings.mlr.press/v70/shrikumar17a.html.

[4] R. High, The era of cognitive systems: An inside look at IBM Watson and how it works, *IBM Corporation, Redbooks* (2012).

[5] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton et al., Mastering the game of go without human knowledge, *Nature* **550**(7676) (2017), 354.

[6] P.W. Koh and P. Liang, Understanding black-box predictions via influence functions, in: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, JMLR. org, 2017, pp. 1885–1894.

[7] T. Berners-Lee, J. Hendler, O. Lassila et al., The semantic web, *Scientific american* **284**(5) (2001), 28–37.

[8] C. Bizer, T. Heath and T. Berners-Lee, Linked data: The story so far, in: *Semantic services, interoperability and web applications: emerging concepts*, IGI Global, 2011, pp. 205–227.

[9] K. Bollacker, C. Evans, P. Paritosh, T. Sturge and J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, in: *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, AcM, 2008, pp. 1247–1250.

[10] S.P. Singh and S. Markovitch (eds), Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA, AAAI Press, 2017. http://www.aaai.org/Library/AAAI/aaai17contents.php.

[11] J. Lang (ed.), Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden, ijcai.org, 2018. ISBN 978-0-9992411-2-7. http://www.ijcai.org/proceedings/2018/.

[12] G.A. Kaminka, M. Fox, P. Bouquet, E. Hüllermeier, V. Dignum, F. Dignum and F. van Harmelen (eds), ECAI 2016 - 22nd European Conference on Artificial Intelligence, 29 August-2 September 2016, The Hague, The Netherlands - Including Prestigious Applications of Artificial Intelligence (PAIS 2016), in *Frontiers in Artificial Intelligence and Applications*, Vol. 285, IOS Press, 2016. ISBN 978-1-61499-671-2.

[13] S.J. Russell and P. Norvig, *Artificial intelligence: a modern approach*, Malaysia; Pearson Education Limited,, 2016.

[14] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg and A. Holzinger, Explainable AI: the new 42?, in: *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, Springer, 2018, pp. 295–303.

[15] M.T. Ribeiro, S. Singh and C. Guestrin, "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, August 13-17, 2016*, 2016, pp. 1135–1144. doi:10.1145/2939672.2939778.

[16] S.M. Lundberg, G.G. Erion and S. Lee, Consistent Individualized Feature Attribution for Tree Ensembles, *CoRR* **abs/1802.03888** (2018). http://arxiv.org/abs/1802.03888.

[17] B. Kim, O. Koyejo and R. Khanna, Examples are not enough, learn to criticize! Criticism for Interpretability, in: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2016, pp. 2280–2288. http://papers.nips.cc/paper/6300-examples-are-not-enough-learn-to-criticize-criticism-for-interpretability.

[18] B.D. Mittelstadt, C. Russell and S. Wachter, Explaining Explanations in AI, in: *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, 2019, pp. 279–288. doi:10.1145/3287560.3287574.

[19] V. Ramanishka, A. Das, J. Zhang and K. Saenko, Top-Down Visual Saliency Guided by Captions, in: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 3135–3144. doi:10.1109/CVPR.2017.334.

[20] M. Craven and J.W. Shavlik, Extracting tree-structured representations of trained networks, in: *Advances in neural information processing systems*, 1996, pp. 24–30.

[21] G.E. Hinton, S. Sabour and N. Frosst, Matrix capsules with EM routing, in: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. https://openreview.net/forum?id=HJWLfGWRb.

[22] Y. Bengio, T. Deleu, N. Rahaman, N.R. Ke, S. Lachapelle, O. Bilaniuk, A. Goyal and C.J. Pal, A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms, *CoRR* **abs/1901.10912** (2019). http://arxiv.org/abs/1901.10912.

[23] J.M.-S. Alexey Ignatiev Nina Narodytska, Abduction-Based Explanations for Machine Learning Models, in: *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, (AAAI-19), Honolulu, Hawaii, USA, 2019*, 2019.

[24] D. Bau, B. Zhou, A. Khosla, A. Oliva and A. Torralba, Network Dissection: Quantifying Interpretability of Deep Visual Representations, in: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 3319–3327. doi:10.1109/CVPR.2017.354.

[25] J. Adebayo, J. Gilmer, M. Muelly, I.J. Goodfellow, M. Hardt and B. Kim, Sanity Checks for Saliency Maps, in: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, 2018, pp. 9525–9536. http://papers.nips.cc/paper/8160-sanity-checks-for-saliency-maps.

[26] B. O'Sullivan, A. Papadopoulos, B. Faltings and P. Pu, Representative Explanations for Over-Constrained Problems, in: *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada*, 2007, pp. 323–328. http://www.aaai.org/Library/AAAI/2007/aaai07-050.php.

[27] U. Junker, QUICKXPLAIN: Preferred Explanations and Relaxations for Over-Constrained Problems, in: *Proceedings of the Nineteenth National Conference on Artificial Intelligence, Sixteenth Conference on Innovative Applications of Artificial Intelligence, July 25-29, 2004, San Jose, California, USA*, 2004, pp. 167–172. http://www.aaai.org/Library/AAAI/2004/aaai04-027.php.

[28] C. Labreuche and S. Fossier, Explaining Multi-Criteria Decision Aiding Models with an Extended Shapley Value, in: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, J. Lang, ed., ijcai.org, 2018, pp. 331–339. ISBN 978-0-9992411-2-7. doi:10.24963/ijcai.2018/46.

[29] L.S. Shapley and M. Shubik, The assignment game I: The core, *International Journal of game theory* **1**(1) (1971), 111–130.

[30] J. Nash, Non-cooperative games, *Annals of mathematics* (1951), 286–295.

[31] L.S. Shapley, A value for n-person games, *Contributions to the Theory of Games* **2**(28) (1953), 307–317.

[32] J. Chen and M.I. Jordan, LS-Tree: Model Interpretation When the Data Are Linguistic, *CoRR* **abs/1902.04187** (2019). http://arxiv.org/abs/1902.04187.

[33] D. Koller and N. Friedman, *Probabilistic Graphical Models - Principles and Techniques*, MIT Press, 2009. ISBN 978-0-262-01319-2. http://mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=11886.

[34] V. Belle, Logic meets Probability: Towards Explainable AI Systems for Uncertain Worlds., in: *IJCAI*, 2017, pp. 5116–5120.

[35] K. Kersting and L. De Raedt, 1 Bayesian Logic Programming: Theory and Tool, *Statistical Relational Learning* (2007), 291.

[36] A. Vellido, J.D. Martín-Guerrero and P.J.G. Lisboa, Making machine learning models interpretable, in: *20th European Symposium on Artificial Neural Networks, ESANN 2012, Bruges, Belgium, April 25-27, 2012*, 2012. https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2012-7.pdf.

[37] S. Rosenthal, S.P. Selvaraj and M.M. Veloso, Verbalization: Narration of Autonomous Robot Experience, in: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, 2016, pp. 862–868. http://www.ijcai.org/Abstract/16/127.

[38] D.J. Brooks, A. Shultz, M. Desai, P. Kovac and H.A. Yanco, Towards State Summarization for Autonomous Robots, in: *Dialog with Robots, Papers from the 2010 AAAI Fall Symposium, Arlington, Virginia, USA, November 11-13, 2010*, 2010. http://www.aaai.org/ocs/index.php/FSS/FSS10/paper/view/2223.

[39] R.K. Sheh, "Why Did You Do That?" Explainable Intelligent Robots, in: *The Workshops of the The Thirty-First AAAI Conference on Artificial Intelligence, Saturday, February 4-9, 2017, San Francisco, California, USA*, 2017. http://aaai.org/ocs/index.php/WS/AAAIW17/paper/view/15162.

[40] K.P. Sycara, M. Paolucci, M.V. Velsen and J.A. Giampapa, The RETSINA MAS Infrastructure, *Autonomous Agents and Multi-Agent Systems* **7**(1–2) (2003), 29–48. doi:10.1023/A:1024172719965.

[41] D. Amir and O. Amir, HIGHLIGHTS: Summarizing Agent Behavior to People, in: *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden, July 10-15, 2018*, 2018, pp. 1168–1176. http://dl.acm.org/citation.cfm?id=3237869.

[42] O. Amir, F. Doshi-Velez and D. Sarne, Agent Strategy Summarization, in: *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden, July 10-15, 2018*, 2018, pp. 1203–1207. http://dl.acm.org/citation.cfm?id=3237877.

[43] W.L. Johnson, Agents that Learn to Explain Themselves., in: *AAAI*, 1994, pp. 1257–1263.

[44] L. Bottou, J. Peters, J. Quiñonero-Candela, D.X. Charles, D.M. Chickering, E. Portugaly, D. Ray, P. Simard and E. Snelson, Counterfactual reasoning and learning systems: The example of computational advertising, *The Journal of Machine Learning Research* **14**(1) (2013), 3207–3260.

[45] A. Aamodt and E. Plaza, Case-based reasoning: Foundational issues, methodological variations, and system approaches, *AI communications* **7**(1) (1994), 39–59.

[46] M. Ghallab, D.S. Nau and P. Traverso, *Automated planning - theory and practice*, Elsevier, 2004. ISBN 978-1-55860-856-6.

[47] M. Fox, D. Long and D. Magazzeni, Explainable Planning, *CoRR* **abs/1709.10256** (2017). http://arxiv.org/abs/1709.10256.

[48] T. Chakraborti, S. Sreedharan, Y. Zhang and S. Kambhampati, Plan Explanations as Model Reconciliation: Moving Beyond Explanation as Soliloquy, in: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, 2017, pp. 156–163. doi:10.24963/ijcai.2017/23.

[49] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick and D. Parikh, Vqa: Visual question answering, in: *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[50] R. Lowe, N. Pow, I. Serban and J. Pineau, The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems, in: *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic*, 2015, pp. 285–294. http://aclweb.org/anthology/W/W15/W15-4640.pdf.

[51] C. Kwok, O. Etzioni and D.S. Weld, Scaling question answering to the web, *ACM Transactions on Information Systems (TOIS)* **19**(3) (2001), 242–262.

[52] H. Kitano and C. Van Ess-Dykema, Toward a plan-based understanding model for mixed-initiative dialogues, in: *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 1991, pp. 25–32.

[53] P.R. Cohen and C.R. Perrault, Elements of a Plan-Based Theory of Speech Acts, in: *Communication in Multiagent Systems, Agent Communication Languages and Conversation Polocies.*, 2003, pp. 1–36. doi:10.1007/978-3-540-44972-0_1.

[54] P.R. Cohen, Back to the Future for Dialogue Research: A Position Paper, *CoRR* **abs/1812.01144** (2018). http://arxiv.org/abs/1812.01144.