Explainable Artificial Intelligence: Concepts, Applications, Research Challenges and Visions

Luca Longo[1(✉)], Randy Goebel[2], Freddy Lecue[3,4], Peter Kieseberg[5], and Andreas Holzinger[2,6]

[1] School of Computer Science, Technological University Dublin, Dublin, Ireland
luca.longo@tudublin.ie
[2] xAI-Lab, Alberta Machine Intelligence Institute, University of Alberta, Edmonton, Canada
rgoebel@ualberta.ca
[3] Inria, Sophia Antipolis, Biot, France
freddy.lecue@inria.fr
[4] Thales, Montreal, Canada
[5] JRC Blockchains, University of Applied Sciences St. Pölten, Sankt Pölten, Austria
Peter.Kieseberg@fhstp.ac.at
[6] Human-Centered AI Lab, Institute for Medical Informatics, Statistics and Documentation, Medical University Graz, Graz, Austria
andreas.holzinger@medunigraz.at

Abstract. The development of theory, frameworks and tools for Explainable AI (XAI) is a very active area of research these days, and articulating any kind of coherence on a vision and challenges is itself a challenge. At least two sometimes complementary and colliding threads have emerged. The first focuses on the development of pragmatic tools for increasing the transparency of automatically learned prediction models, as for instance by deep or reinforcement learning. The second is aimed at anticipating the negative impact of opaque models with the desire to regulate or control impactful consequences of incorrect predictions, especially in sensitive areas like medicine and law. The formulation of methods to augment the construction of predictive models with domain knowledge can provide support for producing human understandable explanations for predictions. This runs in parallel with AI regulatory concerns, like the European Union General Data Protection Regulation, which sets standards for the production of explanations from automated or semi-automated decision making. Despite the fact that all this research activity is the growing acknowledgement that the topic of explainability is essential, it is important to recall that it is also among the oldest fields of computer science. In fact, early AI was re-traceable, interpretable, thus understandable by and explainable to humans. The goal of this research is to articulate the big picture ideas and their role in advancing the development of XAI systems, to acknowledge their historical roots, and to emphasise the biggest challenges to moving forward.

Published by Springer Nature Switzerland AG 2020
A. Holzinger et al. (Eds.): CD-MAKE 2020, LNCS 12279, pp. 1–16, 2020.
https://doi.org/10.1007/978-3-030-57321-8_1

Keywords: Explainable artificial intelligence · Machine learning · Explainability

1 Introduction

Machine learning is often viewed as the technology belonging to the future in many application fields [46], ranging from pure commodities like recommender systems for music, to automatic diagnosis of cancer or control models for autonomous transportation. However, one fundamental issue lies within the realm of explainability [60]. More precisely, most of the existing learning algorithms can often lead to robust and accurate models from data, but in application terms, they fail to provide end-users with descriptions on how they built them, or to produce convincing explanations for their predictions [7]. In many sensitive applications, such as in medicine, law, and other sectors where the main workers are not computer scientists or engineers, the direct application of these learning algorithms and complex models, without human oversight, is currently inappropriate. The reasons are not only technical, like the accuracy of a model, its stability to decisions and susceptibility to attacks, but often arise from sociological concerns, practically settling on the issue of trust. In fact, one of the principal reasons to produce an explanation is to gain the trust of users [13]. Trust is the main way to enhance the confidence of users with a system [66] as well as their comfort while using and governing it [41]. Trust connects to ethics and the intensity of regulatory activities, as for instance the General Data Protection Regulation in the European Union, leads to many legal and even ethical questions: responsibility for safety, liability for malfunction, and tradeoffs therein must inform decision makers at the highest level. Many methods of explainability for data-driven models have emerged in the years, at a growing rate. On the one hand, a large body of work have focused on building post-hoc methods mainly aimed at wrapping fully trained models, often referred to black-boxes, with an explainability layer [37]. A smaller body of research works, on the other hand, have concentrated on creating self-explainable and interpretable models by incorporating explainability mechanisms during their training, often referred to as the ante-hoc phase [7]. Despite the fact that all this research activity is the growing acknowledgement of the topic of explainability [68], by now referred to as Explainable Artificial Intelligence (XAI) [54], it is important to recall that it is also among the oldest fields of computer science. In fact, early AI was retraceable, interpretable, thus understandable by and explainable to humans. For these reasons, many scholars have tried to review research works in the field [1,3,22,51,72]. These reviews reveals the needs for a variety of kinds of explanation, for the identification of methods for explainability and their evaluation as well as the need to calibrate the tradeoffs in the degree or level of explanation appropriate for a broad spectrum of applications.

The goal of this research is to articulate the big picture ideas and their role in advancing the development of XAI systems, to acknowledge their historical roots, and to emphasise the biggest challenges to moving forward. The reminder

of the paper focuses on relevant notions and concepts for explainability in Sect. 2. It then continues in Sect. 3 with descriptions on the applications of methods for XAI and on domains and areas in which these can have a significant impact. A discussion on the research challenges surrounding XAI is presented in Sect. 4. Eventually, recommendations and visions follow by presenting what we believe scholars should focus on in the development of future explainable AI systems.

2 Notions and Related Concepts

A serious challenge for any attempt to articulate the current concepts for XAI is that there is a very high volume of current activity, both on the research side [22,66,72], and in aggressive industrial developments, where any XAI functions can provide a market advantage to all for profit applications of AI [23,60]. In addition, there remains a lack of consensus on terminology, for example as noted within, there are a variety of definitions for the concept of *interpretation*, but little current connection to the formal history, that means in formal theories of explanation or causation [30]. One recent paper [4] provides an organizing framework based on comparing levels of explanation with levels of autonomous driving. The goal is to identify foundational XAI concepts like relationships to historical work on explanation, especially scientific ones, or the importance of interactive explanation as well as the challenge of their evaluation. Note further that the need for a complex system to provide explanations of activities, including predictions, is not limited to those with components created by machine learning (example in [53]). Pragmatically, the abstract identification of a scientific explanation that enables an explainee to recreate an experiment or prediction can arise in very simple circumstances. For example, one can evaluate an explanation by simply noting whether it is sufficient to achieve an explainee's intended task. For example, in Fig. 1, the pragmatic value of an Ikea visual assembly "explanation" is whether the assembler explainee can achieve the assembly using the diagram.

Overall and within this broad spectrum of ideas related to explanation, there is some focus on the foundational connection between explanation and that of abductive reasoning. For example, the historical notion of *scientific explanation* has been the subject of much debate in the community of science and philosophy [70]. Some propose that a theory of explanation should include both scientific and other simpler forms of explanation. Consequently, it has been a common goal to formulate principles that can confirm an explanation as a scientific one. Aristotle is generally considered to be the first philosopher to articulate an opinion that knowledge becomes scientific when it tries to explain the causes of "why." His view urges that science should not only keep facts, but also describe them in an appropriate explanatory framework [15]. In addition to this theoretical view, empiricists also maintain a belief that the components of ideas should be acquired from perceptions with which humans become familiar through sensory experience. The development of the principles of scientific explanation from this perspective prospered with the so-called Deductive-Nomological (DN) model that was described by Hempel in [24–26], and by Hempel and Oppenheim in [27].

There is a more pragmatic AI historical research thread that connects scientific explanation to AI implementations of abductive reasoning. One such thread, among many, begins with Pople in 1973 [59], Poole et al. in 1987 [58], Muggleton in 1991 [52], to Evans et al. in 2018 [14]. Pople described an algorithm for abduction applied to medical diagnosis. Poole et al. provided an extension to first order logic which could subsume non-monotonic reasoning theories and also identify explanatory hypothesis for any application domain. Muggleton proposed a further refinement referred to as inductive logic programming where hypotheses are identified by inductive constraints within any logic, including higher-order logics. Finally, the adoption of this thread of reasoning have been generalised to explanation based on inductive logic programming by Evans et al. [14]. This most recent work connects with information theoretic ideas used to compare differences in how to learn probability distributions that are modeled by machine learning methods.

Interpreting and explaining a model trained from data by employing a machine learning technique is not an easy task. A body of literature has focused on tackling this by attempting at defining the concept of *interpretability*. This has lead to the formation of many types of explanation, with several attributes and structures. For example, it seems to human nature to assign causal attribution of events [23], and we possess an innate psychological tendency to anthropomorphism. As a consequence, an AI-based system that purports to capture *causal relations* should be capable of providing a causal explanation of its inferential process (example in [55]). Causality can been considered a fundamental attribute of explainability, especially when scientific explainability carries a responsibility to help the explainee reconstruct the inferential process leading to a prediction. Many have noted this role on how explanations should make the causal relationships between the inputs and the outputs of a model explicit [17,30,41,51].

Despite the fact that data-driven models are extremely good at discovering associations in the data, unfortunately they can not guarantee causality of these associations. The objective of significantly inferring causal relationships depends on prior knowledge, and very often some of the discovered associations might be completely unexpected, not interpretable nor explainable. As pointed by [1], the decisions taken considering the output of a model should be clearly explainable to support their *justifiability*. These explanations should allow the identification of potential flows both in a model, enhancing its *transparency*, the knowledge *discovery* process, supporting its *controllability* and *improvement* of its accuracy. Although the importance of explainability is clear, the definition of objective criteria to evaluate methods for XAI and validate their explanations is still lacking. Numerous notions underlying the effectiveness of explanations were identified from the fields of Philosophy, Psychology and Cognitive Science. These were related to the way humans define, generate, select, evaluate and present explanations [50].

3 Applications and Impact Areas

Explainable artificial intelligence has produced many methods so far and it has been applied in many domains, with different expected impacts [21]. In these applications, the production of explanations for black box predictions requires a companion method to extract or lift correlative structures from deep-learned models into vocabularies appropriate for user level explanations. Initial activities focused on deep learning image classification with explanations emerging as heat maps created on the basis of gaps in probability distributions between a learned model and an incorrect prediction [5]. However, the field has become so diverse in methods, often determined by domain specific issues and attributes, that it is scarcely possible to get in-depth knowledge on the whole of it. Additionally, one major aspect though is the problem of explainable AI, where lot of problems have been emerged and illustrated in the literature, especially from not being able to provide explanations. While all of these topics require long and in-depth discussions and are certainly of significant importance for the future of several AI methods in many application domains, we want to focus on the benefits that can be reaped from explainability. This means not focusing on the issues of incomplete and imperfect technologies as a stopping point for applications, but discussing novel solutions provided by explainable AI. A discussion of some, partly prominent and partly surprising examples follows, with arguments on why a certain amount of explainability - as a reflection - is required for more advanced AI. There are many sectors that already have fully functional applications based on machine learning, but still serious problems in applying them exist. These are often caused by failing to be capable to explain how these methods work. In other words, it is known that they work, but the concrete results cannot be explained. Many of these applications either come from safety critical or personally sensitive domains, thus a lot of attention is put on explanations of the inferences of trained models, usually predictions or classifications.

Threat Detection and Triage - The detection of threats and efficient triage have been core topics in the area of IT-Security for at least the past three decades. This started with research in the area of code analysis and signature based AntiVirus-Software, moving towards automated decompilation and code analysis, as well as supporting the automated analysis of network monitoring information for triage. Currently, fully automated threat detection and triage is not available in real life systems due to the complexity of the task and the problem with false positives, even though several different approaches exist. These also include strategies that do not try to detect actual threats, but rather filtering out all known legit network travel and thus drastically reducing the amount of information requiring manual analysis [56]. Still, a major problem without explainability lies in the opaque nature of these methods, thus not being able to fully understand their inner functioning and how an inference was reached. Explainability could greatly enhance the detection capabilities, especially since dynamic effects, such as changing user behavior, could be modelled and introduced earlier into the algorithms without generating a large set of false positives.

Explainable Object Detection - Object detection is usually performed from a large portfolio of artificial neural networks (ANN) architectures such as YOLO, trained on large amount of labelled data. In such contexts, explaining object detections is rather difficult if not impossible due to the high complexity of the hyperparameters (number of layers, filters, regularisers, optimizer, loss function) of the most accurate ANNs. Therefore, explanations of an object detection task are limited to features involved in the data and modeled in the form of saliency maps [11] or at best to examples [40], or prototypes [35]. They are the state-of-the-art approaches but explanations are limited by data frames feeding the ANNs. Industrial applications embedding object detection, such as obstacles detection for trains, do require human-like rational for ensuring the system can be guaranteed, even certified [39].

Protection Against Adversarial ML - In adversarial machine learning, attackers try to manipulate the results of learning algorithms by inserting specifically crafted data in the learning process [32], in order to lead a model to learn erroneous things. Detection of such a manipulation is not trivial, especially in contexts with big data, where no model exists before the analysis phase. While there are several proposals on how to deal with this issue [16], some of them employ neural sub-networks for differentiating between malicious and benign input data like [49]. In this specific circumstance, explainability would have a great impact as it will support the task of uncovering such a manipulation far more quickly, efficiently and without actually finding the examples that have been manipulated, thus greatly enhancing trust in machine learning inferences [31].

Open Source Intelligence (OSINT) - In Open Source Intelligence [19], information retrieval is purely reduced to openly available information, as contrary to Signals Intelligence (SIGINT). However, there are several major issues surrounding OSINT, especially referring to context, languages and the amount of information available. Similarly, another problem lies in deciding how much a source is trusted, and what level of impact news of sources shall have on the result of their aggregations. This is especially important when considering adversarial attacks against OSINT methods and systems [12]. Explainability could provide means for detecting these attacks, with an impact on mitigating their influence. Furthermore, the information that an attack against an intelligent system was launched is also a valuable input from an intelligence perspective, so explainability might lead to additional valuable information. However, not all false information exists due to malice, especially when reporting very recent events: information particles might be wrong, misleading or simply unknown at the time of reporting. OSINT becomes especially complex in case of ongoing events, where facts change every minute, either due to knew intelligence, or simply because of changes in the event itself. Explainability would allow to estimate the effects of incorrect information particles on the overall machine learning outcomes, thus allowing, for instance, to give error margins on reported numbers.

Trustworthy (autonomous) Medical Agents - Several architectures for integrating machine learning into medical decision making have been devised in the past. These are based upon a doctor-in-the-loop approach whereby doctors act as input providers to machine learning algorithms. These can lead to suggestions related to diagnosis or treatment that can be subsequently reviewed by the doctors themselves, who, in turn, can provide feedback in a loop to further enhance modeling [34]. Additionally, the mechanism can also introduces external knowledge to support decision making aimed at incorporating the latest findings in the underlying medical field.

Autonomous Vehicles - While certainly being developed within machine learning, explainability would be beneficial for the area of autonomous vehicles, especially considering autonomous cars. In cases of car accidents, explanations can help trace the reasons why an autonomous vehicle behaved in a certain why and took certain actions. Consequently this can not only lead to safer vehicles, but it also can help solve issues in court faster, greatly enhancing trust towards these novel ML-based technologies and especially the resulting artifacts [20].

4 Research Challenges

A number of research challenges surrounding the development of methods for explainability exist, including technical, legal and practical challenges.

4.1 Technical Challenges

XAI Systems Evaluation. The comprehensive study of what explanation means from a sociological viewpoint [50] begs a difficult issue that is both technical and non-technical: *how does one evaluate the quality of an explanation?* It is not a surprise that the quality or value of an explanation is at least partly determined by the receiver of an explanation, sometimes referred to as the "explainee". An easy way to frame the challenge of evaluating explanations, with respect to an explainee, arises from observing the history of the development of evaluation techniques from the field of data visualization [36]. A simple example of "visual explanation" can frame the general evaluation problem for all explanations as follows. Consider the IKEA assembly diagram, rendered in Fig. 1. A simple list of requirements to assess explanation quality emerges from considering the IKEA assembly instructions as a visual explanation of how to assemble the piece of furniture. In this case, the visual explanation is intended to guide all explainees, and not just a single individual, to the successful assembly of the furniture item. One measure of quality is simply to test whether any individual explainee can use the visual explanation to complete the assembly. Another measure is about whether the visual explanation is clear and unambiguous, so that the assembly is time efficient. In the case of Fig. 1, the sequencing of steps might be misinterpreted by an explainee, and that the simple use of circular arrows to indicate motion may also be ambiguous.

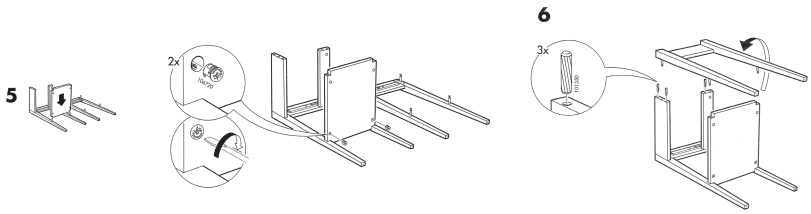

Fig. 1. An IKEA visual explanation for furniture assembly

Overall, and as anticipated in the general evaluation of explanation systems, one could design cognitive experiments to determine, over an experimental human cohort, which portions of the explanation clearly lead to correct inferences, and those which are more difficult to correctly understand. This means that XAI system requirements should include the need to produce an explicit representation of all the components in a way that supports the appropriate interpretation of the visual classification of components. One can generalize visual explanations to the full repertoire that might obtain for a general XAI system. This means a set of representations of the semantics of an underlying domain of application that can provide support to construct an explanation that is understood by a human explainee.

XAI Interpretation. Even though XAI systems are supposed to expose the functioning of a learning technique as well as a set of justification of a model's inferences, it remains rather difficult for a human to interpret them. Explanations are not the final words of an intelligent system but rather the intermediate layer that requires knowledge expertise, context and common-sense characterization for appropriate and correct human interpretation and decision-making [44,64]. Semantics, knowledge graphs [38] and their machine learning representations [6] or similar technical advancements are interesting avenues to be considered for pushing the interpretation at the next right level of knowledge expertise. These might also include the addition of argumentative capabilities, as applied in [45,61] to produce rational and justifiable explanations [62].

4.2 Legal Challenges

While the theoretical ground work in AI stays on the very theoretical side and is thus typically considered to be not problematic from a legal point of view, the actual application of XAI methods in a certain domain can have serious legal implications. This is especially important when considering working with sensitive information. Here, it has yet to be researched whether the explainability related to a model might be used to infer information about individuals, for instance, by using it with slightly different data sets. This technique has been used in many variations in IT-Security, especially considering anonymized data

sets or partially released sensitive information as a basis to gather more intelligence on the people involved [9]. Similar attacks have already been proposed and carried out against machine learned models [65] and these allowed to produce a great amount of information, hidden correlations and causalities that were used to infer sensitive information.

Concepts like federated machine learning are built on the notion of executing machine learning algorithms locally on sensitive data sets and then exchanging the resulting feature sets in order to be combined centrally. These are in contrast to more traditional approaches that collect all sensitive data centrally and then run the learning algorithms. One challenge for federated machine learning is to achieve model robustness but greatly focus on protective sensitive inferences. This justifies the need for more applicable anonymisation techniques, as many of the current methods are unsuitable for many application scenarios, either due to performance or quality issues [47,48]. In addition, other legal challenges exist such as the right to be forgotten [67]. This 'reflects the claim of an individual to have certain data deleted so that third persons can no longer trace them'. This fair right is accompanied by technical difficulties ranging from the issue related to the deletion of entries in modern systems, to the problem of inferring information on individuals from aggregates and especially the removal of said individuals from the aggregation process.

Despite the aforementioned challenges, positive benefits can be brought by explainability to the area of machine learning and AI as a whole with respect to legal issues. While the issue of transparency, a key requirement in the General Data Protection Regulation (GDPR), can be a rather a hard issue to tackle, this could change with explainability providing detailed insight, where, when and to what extent personal data of a single individual was involved in a data analysis workflow [69]. While this is currently not a binding requirement to provide that level of details [69], this could be a game changer regarding acceptance of AI, as well as increasing privacy protection in a data driven society. Furthermore, a significant problem currently tackled in machine learning is bias [71], especially since simple methods for tackling the issue have shown to be ineffective [33]. Explainability could support this combat and thus provide a better legal standing for the results derived from data driven systems, especially when used for socio-economic purposes.

4.3 Practical Challenges

One of the most crucial success factors of AI generally and XAI specifically, is to ensure effective human-AI interfaces to enable a usable and useful interaction between humans and AI [2]. Such goals have been discussed in the HCI community for decades [10], but it was not really seen as important in the AI community. Now the needs and demands of XAI for 'explainable user interfaces' may finally stimulate to realise advanced human-centered concepts similar to the early visions of Vannevar Bush in 1945 [8]. Here, the goal is to explore both the explainability side, that means the artificial explanation generated by machines, as well as the human side, that means the human understanding. In an ideal

world, both machine explanations and human understanding would be identical, and congruent with the ground truth, which is defined for both machines and humans equally. However, in the real world we face two significant problems:

- the ground truth cannot always be fully defined, as for instance when concerned with medical diagnoses [57] when there is high uncertainty;
- human models such as scientific, world, problem solving models, are often based on causality, in the sense of Judea Pearl [55], which is very challenging as current machine learning does not incorporate them and simply follows pure correlation.

Practically speaking, current XAI methods mainly focus on highlighting input–relevant parts, for example via heat-mapping, that significantly contributed to a certain output, or the most relevant features of a training data set that influenced the most the model accuracy. Unfortunately, they do not incorporate the notion of human model, and therefore there is a need to take also into account the concept of causability [30]. In detail, in line with the concept of usability [28], causability is defined as 'the extent to which an explanation of a statement to a human expert achieves a specified level of causal understanding with effectiveness, efficiency and satisfaction in a specified context of use and the particular contextual understanding capabilities of a human'. Following this concept, it becomes possible to measure the quality of explanations in the same terms as usability (effectiveness, efficiency and satisfaction in a specified context of use), for example with a measurement scale [29].

5 Recommendations and Visions

Machine learning, as a solid research area within artificial intelligence, has undoubtedly impacted the field by providing scholars with a robust suite of methods for modeling complex, non-linear phenomena. With the growing body of work in the last decade on deep learning, this impact has significantly expanded to many applications areas. However, despite the widely acknowledged capability of machine and deep learning to allow scholars to induce accurate models from data and extract relevant patterns, accelerating scientific discovery [18], there is the problem of their interpretability and explainability. For this reason, the last few years have seen a growing body of work on research in methods aimed at explaining the inner functioning of data-driven models and the learning techniques used to induce them. Currently and generally recognised as a core area of AI, eXplainable Artificial Intelligence (XAI) has produced a plethora of methods for model interpretability and explainability. Hundred of scientific articles are published each month in many workshops, conferences and presented at symposium around the world. Some of them focus on wrapping trained models with explanatory layers, such as knowledge graphs [39]. Other try to embed the concept of explainability during training, and some of them try to merge learning capabilities with symbolic reasoning [43]. Explainability is a concept borrowed

from psychology, since it is strictly connected to humans, that is difficult to operationalise. A precise formalisation of the construct of explainability is far from being a trivial task as multiple attributes can participate in its definition [61]. Similarly, the attributes might interact with each other, adding complexity in the definition of an objective measure of explainability [42,63]. For these reasons, the last few years have seen also a growing body of research on approaches for evaluating XAI methods. In other words, approaches that are more focused on the explanations generated by XAI solutions, their structure, efficiency, efficacy and impact on humans understanding.

The first recommendation to scholars willing to perform scientific research on explainable artificial intelligence and create XAI methods is to firstly focus on the structure of explanations, the attributes of explainability and the way they can influence humans. This links computer science with psychology. The second recommendation is to define the context of explanations, taking into consideration the underlying domain of application, who they will serve and how. Ultimately, explanations are effective when they help end-users to build a complete and correct mental representation of the inferential process of a given data-driven model. Work on this direction should also focus on which type of explanation can be provided to end-users, including textual, visual, numerical, rules-based or mixed solutions. This links computer science with the behavioural and social sciences. The third recommendation is to clearly define the scope of explanations. This might involve the creation of a method that provide end-users with a suite of local explanations for each input instance or the formation of a method that focuses more on generating explanations on a global level aimed at understanding a model as a whole. This links computer science to statistics and mathematics. The final recommendation is to involve humans, as ultimate users of XAI methods, within the loop of model creation, exploitation, as well as the enhancement of its interpretability and explainability. This can include the development of interactive interfaces that allow end-users to navigate through models, understanding their inner logic at a local or global level, for existing or new input instances. This links artificial intelligence with human-computer interaction.

The visions behind explainable artificial intelligence are certainly numerous. Probably the most important is the creation of models with high accuracy as well as high explainability. The trade-off between these two sides is well known, and usually, increments in one dimension means decrements in the other dimension. Creating interpretable and explainable models that are also highly accurate is the ideal scenario, but since this has been demonstrated to be a hard problem with currents methods of learning and explainability, further research is needed. One possible solution is the creation of models that are fully transparent at all stages of model formation, exploitation and exploration and that are capable of providing local and global explanations. This leads to another vision, which is the use of methods that embed learning capabilities and symbolic reasoning. The former is aimed at generating models and representations with high accuracy for predictive and forecasting purposes, while the latter to explain these

representations in highly interpretable natural language terms, aligned to the way human understand and reason.

Acknowledgements. R.Goebel would like to acknowledge the support of the Alberta Machine Intelligence Institute, which is one of the three Pan Canadian AI Centres. A.Holzinger would like to acknowledge the support of the Austrian Science Fund (FWF), Project: P-32554 "Explainable AI - A reference model of explainable Artificial Intelligence for the Medical Domain".

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)
2. Amershi, S., et al.: Guidelines for human-AI interaction. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM (2019)
3. Arras, L., Osman, A., Müller, K.R., Samek, W.: Evaluating recurrent neural network explanations. In: Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Florence, Italy, pp. 113–126. Association for Computational Linguistics (2019)
4. Atakishiyev, S., et al.: A multi-component framework for the analysis and design of explainable artificial intelligence. (arXiv:2005.01908v1 [cs.AI]) (2020)
5. Babiker, H.K.B., Goebel, R.: An introduction to deep visual explanation. In: NIPS 2017 - Workshop Interpreting, Explaining and Visualizing Deep Learning (2017)
6. Bianchi, F., Rossiello, G., Costabello, L., Palmonari, M., Minervini, P.: Knowledge graph embeddings and explainable AI. CoRR, abs/2004.14843 (2020)
7. Biran, O., Cotton, C.: Explanation and justification in machine learning: a survey. In: IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI), Melbourne, Australia, pp. 8–13. International Joint Conferences on Artificial Intelligence Inc. (2017)
8. Bush, V.: As we may think. Atl. Mon. **176**(1), 101–108 (1945)
9. Cai, Z., He, Z., Guan, X., Li, Y.: Collective data-sanitization for preventing sensitive information inference attacks in social networks. IEEE Trans. Depend. Secure Comput. **15**(4), 577–590 (2016)
10. Card, S.K., Moran, T.P., Newell, A.: Psychol. Hum. Comput. Interact. Erlbaum, Hillsdale (NJ) (1983)
11. Chang, C.-H., Creager, E., Goldenberg, A., Duvenaud, D.: Interpreting neural network classifications with variational dropout saliency maps. Proc. NIPS **1**(2), 1–9 (2017)
12. Devine, S.M., Bastian, N.D.: Intelligent systems design for malware classification under adversarial conditions. arXiv preprint, arXiv:1907.03149 (2019)
13. Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., Beck, H.P.: The role of trust in automation reliance. Int. J. hum. Comput. Stud. **58**(6), 697–718 (2003)
14. Evans, R., Greffenstette, E.: Learning explanatory rules from noisy data. J. Artif. Intell. Res. **61**, 1–64 (2018)
15. Falcon, A.: Aristotle on causality. Stanford Encyclopedia of Philosophy (2006). (https://plato.stanford.edu
16. Feinman, R., Curtin, R.R., Shintre, S., Gardner, A.B.: Detecting adversarial samples from artifacts. arXiv preprint, arXiv:1703.00410 (2017)

17. Fox, M., Long, D., Magazzeni, D.: Explainable planning. In: IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI), Melbourne, Australia, pp. 24–30. International Joint Conferences on Artificial Intelligence Inc (2017)
18. Gil, Y., Greaves, M., Hendler, J., Hirsh, H.: Amplify scientific discovery with artificial intelligence. Science **346**(6206), 171–172 (2014)
19. Glassman, M., Kang, M.J.: Intelligence in the internet age: the emergence and evolution of open source intelligence (OSINT). Comput. Hum. Behav. **28**(2), 673–682 (2012)
20. Glomsrud, J.A., Ødegårdstuen, A., Clair, A.L.S., Smogeli, Ø.: Trustworthy versus explainable AI in autonomous vessels. In: Proceedings of the International Seminar on Safety and Security of Autonomous Vessels (ISSAV) and European STAMP Workshop and Conference (ESWC) 2019, pp. 37–47. Sciendo (2020)
21. Goebal, R., et al.: Explainable AI: the new 42? In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) CD-MAKE 2018. LNCS, vol. 11015, pp. 295–303. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99740-7_21
22. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. (CSUR) **51**(5), 93:1–93:42 (2018)
23. Ha, T., Lee, S., Kim, S.: Designing explainability of an artificial intelligence system. In: Proceedings of the Technology, Mind, and Society, p. 1, article no. 14, Washington, District of Columbia, USA. ACM (2018)
24. Hempel, C.G.: The function of general laws in history. J. Philos. **39**(2), 35–48 (1942)
25. Hempel, C.G.: The theoretician's dilemma: a study in the logic of theory construction. Minnesota Stud. Philos. Sci. **2**, 173–226 (1958)
26. Hempel, C.G.: Aspects of Scientific Explanation. Free Press, New York (1965)
27. Hempel, C.G., Oppenheim, P.: Studies in the logic of explanation. Philos. Sci. **15**(2), 135–175 (1948)
28. Holzinger, A.: Usability engineering methods for software developers. Commun. ACM **48**(1), 71–74 (2005)
29. Holzinger, A., Carrington, A., Müller, H.: Measuring the quality of explanations: the System Causability Scale (SCS). KI - Künstliche Intelligenz **34**(2), 193–198 (2020). https://doi.org/10.1007/s13218-020-00636-z
30. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Mueller, H.: Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. **9**(4), e1312 (2019)
31. Holzinger, K., Mak, K., Kieseberg, P., Holzinger, A.: Can we trust machine learning results? Artificial intelligence in safety-critical decision support. ERCIM NEWS **112**, 42–43 (2018)
32. Huang, L., Joseph, A.D., Nelson, B., Rubinstein, B.I., Tygar, J.D.: Adversarial machine learning. In: Proceedings of the 4th ACM workshop on Security and artificial intelligence, pp. 43–58 (2011)
33. Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Fairness-aware classifier with prejudice remover regularizer. In: Flach, P.A., De Bie, T., Cristianini, N. (eds.) ECML PKDD 2012. LNCS (LNAI), vol. 7524, pp. 35–50. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33486-3_3
34. Kieseberg, P., Malle, B., Frühwirt, P., Weippl, E., Holzinger, A.: A tamper-proof audit and control system for the doctor in the loop. Brain Inform. **3**(4), 269–279 (2016). https://doi.org/10.1007/s40708-016-0046-2

35. Kim, B., Koyejo, O., Khanna, R.: 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, Barcelona, Spain, 5–10 December, pp. 2280–2288 (2016)
36. Lam, H., Bertini, E., Isenberg, P., Plaisant, C., Carpendale, S.: Empirical studies in information visualization: seven scenarios. IEEE Trans. Graph. Vis. Comput. **18**(9), 1520–1536 (2012)
37. Laugel, T., Lesot, M.J., Marsala, C., Renard, X., Detyniecki, M.: The dangers of post-hoc interpretability: unjustified counterfactual explanations. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, (IJCAI), Macao, China, pp. 2801–2807 2019. International Joint Conferences on Artificial Intelligence Organization (2019)
38. Lécué, F.: On the role of knowledge graphs in explainable AI. Semant. Web **11**(1), 41–51 (2020)
39. Lécué, F., Pommellet, T.: Feeding machine learning with knowledge graphs for explainable object detection. In: Suárez-Figueroa, M.C., Cheng, G., Gentile, A.L., Guéret, C., Keet, C.M., Bernstein, A., (eds.) Proceedings of the ISWC 2019 Satellite Tracks (Posters & Demonstrations, Industry, and Outrageous Ideas) co-located with 18th International Semantic Web Conference (ISWC 2019), 26–30 October 2019, Auckland, New Zealand, volume 2456 of CEUR Workshop Proceedings, pp. 277–280. CEUR-WS.org (2019)
40. Li, O., Liu, H., Chen, C., Rudin, C.: Deep learning for case-based reasoning through prototypes: a neural network that explains its predictions. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), 2–7 February 2018, New Orleans, Louisiana, USA, pp. 3530–3537 (2018)
41. Lipton, Z.C.: The mythos of model interpretability. Commun. ACM **61**(10), 36–43 (2018)
42. Longo, L.: Argumentation for knowledge representation, conflict resolution, defeasible inference and its integration with machine learning. In: Holzinger, A. (ed.) Machine Learning for Health Informatics. LNCS (LNAI), vol. 9605, pp. 183–208. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-50478-0_9
43. Longo, L., Dondio, P.: Defeasible reasoning and argument-based systems in medical fields: an informal overview. In: 2014 IEEE 27th International Symposium on Computer-Based Medical Systems, pp. 376–381. IEEE (2014)
44. Longo, L., Hederman, L.: Argumentation theory for decision support in healthcare: a comparison with machine learning. In: Imamura, K., Usui, S., Shirao, T., Kasamatsu, T., Schwabe, L., Zhong, N. (eds.) BHI 2013. LNCS (LNAI), vol. 8211, pp. 168–180. Springer, Cham (2013). https://doi.org/10.1007/978-3-319-02753-1_17
45. Longo, L., Kane, B., Hederman, L.: Argumentation theory in health care. In: 2012 25th International Symposium on Computer-Based Medical Systems (CBMS), pp. 1–6. IEEE (2012)
46. Makridakis, S.: The forthcoming artificial intelligence (AI) revolution: its impact on society and firms. Futures **90**, 46–60 (2017)
47. Malle, B., Kieseberg, P., Holzinger, A.: Do not disturb? Classifier behavior on perturbed datasets. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) CD-MAKE 2017. LNCS, vol. 10410, pp. 155–173. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66808-6_11

48. Malle, B., Kieseberg, P., Weippl, E., Holzinger, A.: The right to be forgotten: towards machine learning on perturbed knowledge bases. In: Buccafurri, F., Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) CD-ARES 2016. LNCS, vol. 9817, pp. 251–266. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-45507-5_17
49. Metzen, J.H., Genewein, T., Fischer, V., Bischoff, B.: On detecting adversarial perturbations. arXiv preprint, arXiv:1702.04267 (2017)
50. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019)
51. 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. In: IJCAI Workshop on Explainable AI (XAI), Melbourne, Australia, pp. 36–42. International Joint Conferences on Artificial Intelligence Inc. (2017)
52. Muggleton, S.: Inductive logic programming. New Generat. Comput. **8**(4), 295–318 (1991)
53. Nunes, I., Jannach, D.: A systematic review and taxonomy of explanations in decision support and recommender systems. User Model. User Adap. Interact. **27**(3), 393–444 (2017). https://doi.org/10.1007/s11257-017-9195-0
54. Páez, A.: The pragmatic turn in explainable artificial intelligence (XAI). Mind. Mach. **29**, 1–19 (2019)
55. Pearl, J.: Causality: Models, Reasoning, and Inference, 2nd edn. Cambridge University Press, Cambridge (2009)
56. Pirker, M., Kochberger, P., Schwandter, S.: Behavioural comparison of systems for anomaly detection. In: Proceedings of the 13th International Conference on Availability, Reliability and Security, pp. 1–10 (2018)
57. Pohn, B., Kargl, M., Reihs, R., Holzinger, A., Zatloukal, k., Müler, H.: Towards a deeper understanding of how a pathologist makes a diagnosis: visualization of the diagnostic process in histopathology. In: IEEE Symposium on Computers and Communications (ISCC 2019). IEEE (2019)
58. Poole, D., Goebel, R., Aleliunas, R.: Theorist: A logical reasoning system for defaults and diagnosis. The Knowledge Frontier. Symbolic Computation (Artificial Intelligence), pp. 331–352 (1987). https://doi.org/10.1007/978-1-4612-4792-0_13
59. Pople, H.: On the mechanization of abductive logic. In: IJCAI'1973: Proceedings of the 3rd International Joint Conference on Artificial Intelligence, pp. 147–152. Morgan Kaufmann Publishers (1973)
60. Preece, A.: Asking "why" in AI: explainability of intelligent systems-perspectives and challenges. Intell. Syst. Account. Financ. Manage. **25**(2), 63–72 (2018)
61. Rizzo, L., Longo, L.: Inferential models of mental workload with defeasible argumentation and non-monotonic fuzzy reasoning: a comparative study. In: Proceedings of the 2nd Workshop on Advances in Argumentation in Artificial Intelligence, co-located with XVII International Conference of the Italian Association for Artificial Intelligence, AI^3@AI*IA 2018, 20–23 November 2018, Trento, Italy, pp. 11–26 (2018)
62. Rizzo, L., Longo, L.: A qualitative investigation of the explainability of defeasible argumentation and non-monotonic fuzzy reasoning. In: Proceedings for the 26th AIAI Irish Conference on Artificial Intelligence and Cognitive Science Trinity College Dublin, 6–7 December 2018, Dublin, Ireland, pp. 138–149 (2018)
63. Rizzo, L., Longo, L.: An empirical evaluation of the inferential capacity of defeasible argumentation, non-monotonic fuzzy reasoning and expert systems. Expert Syst. Appl. **147**, 113220 (2020)

64. Rizzo, L., Majnaric, L., Longo, L.: A comparative study of defeasible argumentation and non-monotonic fuzzy reasoning for elderly survival prediction using biomarkers. In: Ghidini, C., Magnini, B., Passerini, A., Traverso, P. (eds.) AI*IA 2018. LNCS (LNAI), vol. 11298, pp. 197–209. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-03840-3_15
65. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 3–18. IEEE (2017)
66. Tintarev, N., Masthoff, J.: A survey of explanations in recommender systems. In: IEEE 23rd international conference on data engineering workshop, pp. 801–810, Istanbul, Turkey. IEEE (2007)
67. Villaronga, E.F., Kieseberg, P., Li, T.: Humans forget, machines remember: artificial intelligence and the right to be forgotten. Comput. Law Secur. Rev. **34**(2), 304–313 (2018)
68. Vilone, G., Longo, L.: Explainable artificial intelligence: a systematic review. CoRR, abs/2006.00093 (2020)
69. Wachter, S., Mittelstadt, B., Floridi, L.: Transparent, explainable, and accountable AI for robotics. Sci. Robot. **2**(6) (2017)
70. Woodward, J.: Scientific explanation. Stanford Encyclopedia of Philosophy (2003). (https://plato.stanford.edu
71. Yapo, A., Weiss, J.: Ethical implications of bias in machine learning. In: HICCS 2018, Proceedings of the 51st Hawaii International Conference on System Sciences (2018)
72. Zhang, Q., Zhu, S.: Visual interpretability for deep learning: a survey. Front. Inform. Technol. Electron. Eng. **19**(1), 27–39 (2018). https://doi.org/10.1631/FITEE.1700808