Data Mining and Knowledge Discovery (2024) 38:2571–2595
https://doi.org/10.1007/s10618-024-01041-y

GUEST EDITORIAL

Explainable and interpretable machine learning and data mining

Martin Atzmueller[1,5] · Johannes Fürnkranz[2] · Tomáš Kliegr[3] · Ute Schmid[4]

Accepted: 22 May 2024 / Published online: 30 July 2024

Abstract
The growing number of applications of machine learning and data mining in many domains—from agriculture to business, education, industrial manufacturing, and medicine—gave rise to new requirements for how to inspect and control the learned models. The research domain of explainable artificial intelligence (XAI) has been newly established with a strong focus on methods being applied post-hoc on black-box models. As an alternative, the use of interpretable machine learning methods has been considered—where the learned models are white-box ones. Black-box models can be characterized as representing implicit knowledge—typically resulting from statistical and neural approaches of machine learning, while white-box models are explicit representations of knowledge—typically resulting from rule-learning approaches. In this introduction to the special issue on 'Explainable and Interpretable Machine Learning and Data Mining' we propose to bring together both perspectives, pointing out commonalities and discussing possibilities to integrate them.

Keywords Explainable AI · Explainable and interpretable machine learning · Explainable and interpretable data mining · Hybrid artificial intelligence

1 Introduction

Algorithms are affecting an increasing number of aspects of human life, ranging from the preselection of news items and posts in one's social network feed to high-stake medical decisions, e.g., for recommendation and decision support. In many cases, humans have the option to decide whether to follow the algorithmic advice (or recommendation) and to what extent. Existing research seems to indicate a current tendency towards algorithm appreciation, the phenomenon of assigning more weight to algorithmic advice than from a human source (Logg et al. 2019; Bogert et al. 2021). Public trust brings with it responsibility for data scientists to devise reliable

Martin Atzmueller, Johannes Fürnkranz, Tomáš Kliegr and Ute Schmid have contributed equally to this work.

Extended author information available on the last page of the article

models. This is often hard to achieve due to factors including data quality issues or models getting obsolete in time due to concept shift. To retain and strengthen human trust, it is therefore vital for the machine to explain the reasoning behind its prediction and also to articulate its (un)certainty in a way that can be understood by humans.

In response to these challenges, scientific discourse in artificial intelligence and data science has focused on explainable AI (XAI) (Gunning et al. 2019; Arrieta et al. 2020). This emerging field covers algorithmic transparency, interpretability, accountability and explainability of algorithmic models and decisions (Adadi and Berrada 2018; Molnar 2022; Samek et al. 2019; Rudin 2019; Górriz et al. 2023). In machine learning, data mining and knowledge discovery, the respective approaches can be classified as intrinsically interpretable (white-box) or non-directly interpretable (black-box). White-box models, such as rule learners (Fürnkranz et al. 2012), (local) pattern mining (Morik et al. 2005; Fürnkranz and Knobbe 2010) and inductive logic programming (Cropper et al. 2022), result in explicit models which are inherently interpretable. In contrast, black-box approaches, such as ensembles or (deep) neural networks, result in opaque models. For this second type of models, over the last years, different approaches for ex-post explanation generation have been proposed. Interpretability and explainability are particularly important for knowledge discovery and data mining, where the understandability of the found patterns is a key factor in the classic definition of the field (Fayyad et al. 1996), being relevant both for implementation as well as application.

The goal of this special issue is to provide a joint perspective on work in explainable and interpretable machine learning and data mining. Integrating these areas should enable new perspectives on questions regarding appropriate learning formalisms, interpretation and explanation techniques, their metrics, as well as the respective assessment options that arise. In this introductory article, we do not aim to provide yet another review but want to present and highlight special demands on explainable and interpretable approaches for machine learning and data mining, which are also reflected in the articles contained in the special issue.

In the remainder of this editorial, we will first provide a brief introduction to the basic concepts of explainability and interpretability (Sect. 2). Thereafter, in Sect. 3, we outline current research directions in explainable AI, with a particular focus on data mining and machine learning. The topics covered include measuring interpretability, a survey of types of interpretable models, the role of background knowledge, and emerging topics such as applications involving complex data and combining reasoning and learning. Section 4 provides a brief overview of the articles included in this special issue, before we conclude with an outlook on the future of explainable AI.

2 Basic concepts in explainability and interpretability

Especially in complex and sensitive domains involving high stake decisions, for example, agriculture (Schoenke et al. 2021; Ryo 2022), medicine (Holzinger et al. 2017; Atzmueller et al. 2005), cyber security (Atzmueller et al. 2023; Nadeem et al.

2023), industry (Gade et al. 2019; Ahmed et al. 2022), or law and judicial applications (Deeks 2019; Lim 2021), human agency and oversight are important requirements for the trustworthiness of AI-supported decision making as incorporated in respective decision support systems (Laux et al. 2023; Schmid 2023). In general, it might not be necessary for a system to provide explanations in all possible situations—in fact, this might be even harmful, resulting in cognitive overload (Ai et al. 2021) or reduced efficiency. However, it should be possible to obtain an explanation for a model's decision whenever a human requests an explanation or whenever a system output is assessed to be uncertain. Such explanations should be faithful, that is, being closely related to the model, and at the same time providing the information which are relevant for the current context. An overview of the need for explanations, our working definition for interpretability and explainability, human-centricity of explanations, as well as societal requirements on explanations will be discussed in the following.

2.1 The need for explanations

The models learned by many state-of-the-art machine learning algorithms are often too complex to be directly understood. Such models are referred to as "black boxes". The loss of explainability associated with a technological advance in AI has been foreseen already by Asimov (1950):

> Machines are a gigantic extrapolation. Thus, a team of mathematicians work several years calculating a positronic brain equipped to do certain similar acts of calculation. Using this brain they make further calculations to create a still more complicated brain, which they use again to make one still more complicated and so on. [...] Perhaps roboticists as a whole should now die, since we can no longer understand our own creations.

Fortunately, a number of (admittedly still complicated) algorithms have been developed to provide explanations for the models and predictions generated by the black-box models, including deep neural networks as the current manifestation of Asimov's "positronic brain".

An alternative to explaining a black box is using a directly interpretable model such as a rule list or a decision tree. Even though it is generally believed that using a simpler, more interpretable model may result in a decrease in predictive performance, it has in fact been argued that this accuracy-interpretability tradeoff is a fallacy and that more interpretable models "are more (and not less) accurate" (Rudin and Radin 2019). Nevertheless, while this may be true for a particular setup and data, comprehensive benchmarks have shown that, on average, more complex black-box models such as decision tree ensembles and random forests have a better predictive performance than the typical representatives of the current generation of "white-box" models such as single trees or rule lists (Fernández-Delgado et al. 2014; Freitas 2019).

The current performance impediment does not make research in interpretable models obsolete but quite the opposite, with white-box models being used as

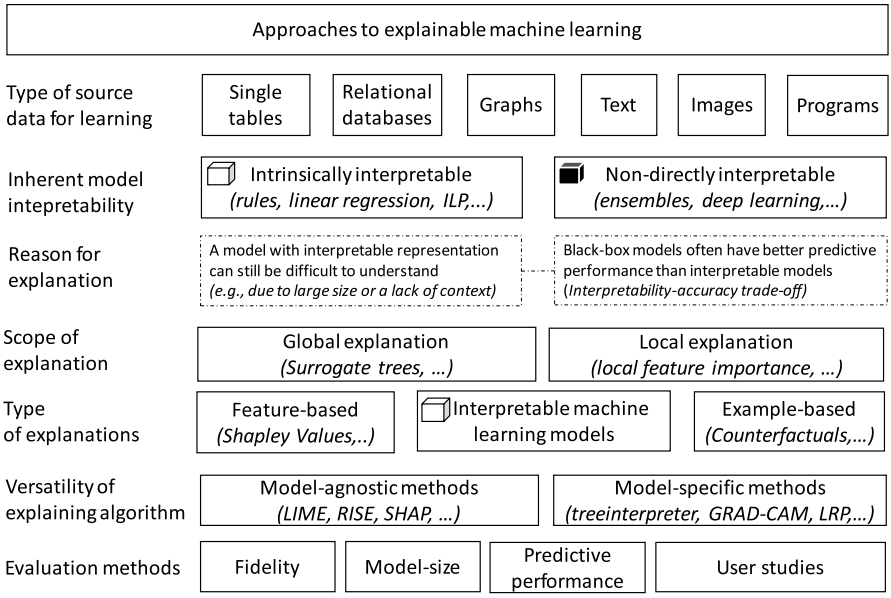

Fig. 1 Interpretable approaches and ex-post XAI methods

elements of ensembles (e.g., decision trees in tree ensembles), for surrogate modelling used to "explain the black box", or for problems where interpretability is more important than the predictive performance. This is also reflected in the call for papers for this special issue which invited articles advancing both explanation of black boxes as well as further refinement of directly interpretable models.

2.2 Taxonomy of explainable machine learning

Figure 1 provides an overview of different approaches or perspectives on explainable machine learning methods.

Type of source data for learning

One of the main factors that determines the choice of an explanation algorithm is the type of data being analyzed, since different methods might be needed for image data as opposed to text. However, versatile methods such as LIME (Ribeiro et al. 2016) can be applied to text, image, and tabular data.

The choice of input data also often determines options for the machine learning algorithm. Tabular data might be amenable to white-box, intrinsically interpretable algorithms such as decision trees or rules. In contrast, achieving good results on image data or texts often requires models with a high number of parameters, such as deep neural networks or ensembles.

Inherent model interpretability and reason for explanations

Rudin (2019) argued very convincingly that whenever it is possible to apply interpretable machine learning, such approaches should be preferred since it is much more straightforward to induce a white-box model directly than to first learn a black-box model and work hard to make it interpretable afterwards. However, this proposition is only feasible for domains and data sets for which interpretable approaches are applicable in such a way that the learned models can have high predictive accuracy. As noted earlier, extensive benchmarks have shown that there is, on average, a notable gap between the performance of the best white-box methods. This shows that there is an *interpretability-accuracy trade-off* between the choice of decision trees, rules and other interpretable models on one hand, and best-performing black-box models such as tree ensembles or neural networks on the other hand.

Interpretable models (mostly) explain themselves.

Interpretable models refer to symbolic models which are inherently interpretable, that is, regression models and rule-based approaches such as decision tree methods, or inductive logic programming. These methods can be often applied to tabular data which are abundant in many domains such as administration, customer relationship, finance, or medical databases.

Explanation generation for black-box models.

These methods include ensembles, (deep) neural networks, and other forms of models, which are difficult if not impossible to interpret for humans already on the syntactical level. For a human to be able to understand and evaluate why a learned black-box model returns a specific output to a given input, an ex-post explanation method is necessary.

Scope of explanations

An important practical question often determined by the given use case is whether it is sufficient to provide an explanation for a specific prediction, or whether the complete (approximated) behavior of a machine learning model should be reviewed.

Local explanations.

This scope of explanation is, to our knowledge, prevalent in the current research. An example is explaining why for a specific input, the model returns a specific output—for instance, classifying an input image as representing a specific type of tumor.

Global explanation.

This scope of explanation aims at explaining the full model. For instance, what information, in general, the model uses for classifying a specific type of tumor. Interpretable models can be seen as global explanations since they are given in a

symbolic form. A common example is a surrogate decision tree. Specialized algorithms include learning near-miss explanations for rule sets via inductive logic programming (Rabold et al. 2022).

Type of explanation

The choice of the explanation algorithm is often determined by how the explanation should look, which might be a critical choice for its information value and, eventually, user acceptance of the explanation.

Feature-based methods.

Many proposed methods analyse feature relevance. That is, they highlight what information in the input data is mostly used to determine the output of the explained model.

Example-based methods and counterfactuals.

Another type of explanation are counterfactuals (Wachter et al. 2017), which is one of the most common types of example-based methods. Here, an explanation shows what needs to be minimally changed in the input to result in a different output for the model. This type of explanation is often discussed in the context of explanations for end-users, for instance, to explain why a loan might be denied. Related to counterfactuals are contrastive (Dhurandhar et al. 2018; Kim et al. 2016) or near-miss (Rabold et al. 2022; Herchenbach et al. 2022) explanations which can be applied to arbitrary data including images.

Explanations through interpretable machine-learning models.

Interpretable approaches have also been proposed as a method to explain opaque machine learning models. For classical feed-forward neural networks, extracting decision trees (TREPAN; Craven and Shavlik 1995) or rules in propositional logic (KBANN; Towell and Shavlik 1994) have been introduced for global explanations. Furthermore, it has been shown that LIME's predominant linear regression method to construct local explanations can be replaced with rule-based (Guidotti et al. 2019) or more expressive relational models (Rabold et al. 2020). Finally, there is research on extracting automata from deep neural networks (Weiss et al. 2018). Such combinations of neural network models with symbolic rules are a special type of hybrid AI (Towell and Shavlik 1994) or neuro-symbolic AI (d'Avila Garcez and Lamb 2023; De Raedt et al. 2020).

Versatility of the explanation algorithm

Another perspective in terms of versatility is whether the explanation algorithm has been designed for a specific type of machine learning method or is model agnostic. In some machine learning workflows, the choice of a learning algorithm is tuned as part of, e.g., an automated machine learning process. Not only in such cases, it

is imperative that the explanation process is not impacted by a swap in the learning algorithm.

Model-specific methods.

Model specific methods exist for various types of the underlying models. For random forests, one approach based on analyzing the tree-based structure of the model has been introduced by Palczewska et al. (2013). A similar idea is used in the python package `treeinterpreter`.[1] Popular methods for neural networks are Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al. 2017) and Layer-wise Relevance Propagation (LRP) (Bach et al. 2015).

Model-agnostic methods.

Alternatively, there are model-agnostic methods where the computation of explanations is based on perturbations or other analytical methods of the input. Well-known methods are LIME (Ribeiro et al. 2016), RISE (Petsiuk et al. 2018), and SHAP (Lundberg and Lee 2017a).

While model-agnostic methods are more generally applicable, it has been suggested that model-specific methods often provide better explanation outcomes. For example, this has been discussed by a recent study showing that LIME is outperformed by Grad-CAM in the task of explaining convolutional neural networks trained to label images (Cian et al. 2020).

Evaluating explanations

Algorithmic explanations can lead to an increase in trust in AI recommendations (Bansal et al. 2021). This result is encouraging but entails responsibility. One could argue that even an imperfect explanation is better than no explanation at all. However, Bansal et al. (2021) also observed that explanations of incorrect AI recommendations increased the chance of the wrong recommendation being accepted by humans. Conversely, AI explanations, even when correct, may reinforce, rather than reverse strong wrong prior human beliefs (Bauer et al. 2023). Especially in high-risk domains, it is imperative to perform a multi-thronged evaluation of explanation algorithms as well as of specific generated explanations.

Fidelity.

The most important requirement for an explanation is that it is faithful to the model, that is, that the explanation really reflects what the model is doing (Lakkaraju et al. 2019). While interpretable methods are inherently faithful, this is not the case for ex-post computations of explanations. For instance, it could be shown that the explanations generated by LIME in an image classification problem are highly dependent on the methods by which superpixels are determined (Schallner et al. 2020).

[1] https://pypi.org/project/treeinterpreter/.

Model size.

The model size is particularly important for interpretable models. Although its syntax might be human-understandable, a particular model might be too complex to be directly understandable to humans. For example, we can consider a decision tree with hundreds of splits or, at the other extreme, a very concise model that further uses non-descriptive feature names. As a result, model size is one of the most common metrics, its main advantage being that it can be computed directly from the training data and even optimized for during the learning of interpretable models. However, it has also been argued that too simple explanations can also be less convincing than more complex ones (Fürnkranz et al. 2020).

Predictive performance.

Predictive metrics such as accuracy or the area under the ROC curve are often relevant for the evaluation of explanation methods. For surrogate models, it is important to know if the simpler yet understandable model performs acceptably well. Predictive performance can also be used to compare explanation algorithms. Using the ROAR approach (RemOve and Retrain) the best explanation algorithm will identify those features as most important whose removal will cause the biggest decrease in model performance relative to other methods (Hooker et al. 2019).

User studies.

The problem is not only ensuring the correctness of the prediction (and explanation) but also whether it will be correctly understood by the target audience, especially given that multiple explanation techniques may be simultaneously required (Reed et al. 2021). As the AI explanation may be out of sync with the quality of the prediction, the goal of XAI should be to accurately *inform* rather than to *convince*; nevertheless, the latter is the objective of many explanation algorithms (Bauer et al. 2023). It is, therefore, important to perform user studies aimed at evaluating the effectiveness of individual explanation methods with correctly stated objectives and evaluation measures.

2.3 Societal requirements on explanations

Until recently, laws governing AI were only in science fiction. The most well-known are the *Three Laws of Robotics* (Asimov 1942), which were—in Asimov's world—hard-wired in robots to prevent them harm human beings, make them obey human orders and preserve their existence.

As of the writing of this editorial, the EU is preparing its Artificial Intelligence Act, which aims to prevent AI from causing harm to humans.[2] One of the means

[2] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

designated to achieve the safety of "high-risk" applications of AI are requirements on transparency and explainability. The draft regulation demands that the overseeing individuals should be able to "correctly interpret the high-risk AI system's output". An open question is which existing XAI algorithms will ensure this? According to an analysis by Reed et al. (2021), current law and regulations expect explanations in the form of a narrative, a story. This is met by explanations provided by Asimov's robots but not by the current generation XAI.

In general, it is argued that XAI supports trustworthiness (Schmid 2023). However, there might be the danger that humans readily accept a model output as trustworthy if some explanation is given, but trust might not be justified. For instance, humans might be biased towards the predictions of a model (Schmidt and Biessmann 2019). Additional criteria have to be met to make a model trustworthy, among them its robustness, fairness and corrigibility (Schmid 2023), also extending to verification and testing (Huang et al. 2020). Furthermore, besides evaluation criteria ensuring the quality of the learned models themselves, the quality of the generated explanations have to fulfill certain criteria, furthermost their faithfulness to the model (Hedström et al. 2023) but also their complexity and comprehensibility to the human (Ai et al. 2021; Göbel et al. 2022).

3 Research directions in explainability and interpretability

The field of XAI has grown considerably in recent years. It is beyond the scope of this editorial to provide a full survey of the field. However, in the following, we nevertheless outline a few important current research directions, including how to measure interpretability (Sect. 3.1), the duality of explaining black-box models as opposed to directly learning white-box models (Sect. 3.2), XAI for complex data types (Sect. 3.3), and the integration of background knowledge into the explanations (Sect. 3.4).

3.1 Measuring interpretability

Obviously, many papers measure interpretability using the complexity of concepts, and whether they can be syntactically read or not. This is debatable, because shorter descriptions are not necessarily more understandable (e.g., textbooks may contain a lot of redundancy). On the other hand, however, too complex models are not interpretable, for example, random forests. In any case, only few papers explicitly define and evaluate interpretability.

Interpretability of model class

Some model classes are considered to be inherently interpretable, including logical models such as decision trees or rules, probabilistic models such as Naïve Bayes classifiers or Bayesian networks, or statistical models such as linear discriminators. As explanations, typically feature weights or heat maps, as well as special

data points such as prototypes, nearest neighbors or counterfactual examples are also considered to be interpretable. The coarsest way of assessing interpretability is to simply state that the learned model or the produced explanation is within one of these classes.

Syntactic properties

Finer grained measures of interpretability often focus on syntactic properties of the model, most notably the complexity of a model. Principles like Occam's Razor or Minimum Description Length (MDL) are commonly used heuristics for model selection, and have shown to be successful in overfitting avoidance. As a consequence, most rule learning algorithms have a strong bias towards simple theories. Nevertheless, the view that more complex models are necessarily less interpretable has recently been questioned as discussed earlier (Sect. 2.2).

Semantic properties

Some works also consider semantic properties of the models, such as how similar the important features of a learned model are to each other or the user's background knowledge. For example, Gabriel et al. (2014) proposed to consider the semantic coherence of its conditions when formulating a rule. Pazzani et al. (2001) show that rules that respect monotonicity constraints are more acceptable to experts than rules that do not.

User acceptance

It has also been argued that interpretability should go beyond the mere ability to understand a learned model. By that view, it is not sufficient that a model can be read and interpreted by its intended user, but that user may assess different models or model classes with different degrees of interpretability. For example, early studies such as (Kononenko 1993; Huysmans et al. 2011; Allahyari and Lavesson 2011; Piltaver et al. 2016) compared different types of machine learning models with respect to their interpretability as perceived by users, or Fürnkranz et al. (2020) report on a crowd-sourcing study where the plausibility of various learned rule-bases was assessed by human subjects.

Utility

Only rarely it has been attempted to directly measure the utility of the learned model or the provided explanation to the user. One of the few attempts for such an operational definition of interpretability is given by Schmid et al. (2017) and Muggleton et al. (2018), who tested whether participants in their study can successfully apply the acquired knowledge to new problems, and thus related interpretability to objective measurements such as the time needed for inspecting a learned concept, for applying it in practice, or for giving it a meaningful and correct name.

Trustworthiness

As already mentioned in Sect. 2.3, trustworthiness is an important factor which can be supported by interpretable and explainable methods and approaches. As stated by Li et al. (2022), for example, "an interpretation algorithm is trustworthy if

it properly reveals the underlying rationale of a model making decisions." In particular, if the respective algorithmic mechanisms of a model towards its decision making are understandable, then the transparency of the model is positively impacted (Ali et al. 2023). Therefore, trustworthiness includes many aspects discussed above, in particular also extending to robustness, transparency, and evaluation of explanations in general (Hedström et al. 2023; Kaur et al. 2021; Vilone and Longo 2021; Schoenherr et al. 2023).

Quantitatively, the frequency of use of the various techniques in the literature seems to be diminishing in roughly the above order. Syntactic assessments of interpretability can be more frequently found than semantic assessments, which in turn are still more common than actual user studies. Arguably, the main reason for this reduction is that these measures become increasingly more difficult and elaborate to obtain. For syntactic measures, one only needs to look at the model class, and possibly compute some obvious features such as the model complexity. These can thus be easily employed in empirical studies, in much the same way as predictive measures such as accuracy can be computed without much effort. Semantic measures are harder to obtain, as they would, e.g., require additional background knowledge in the form of knowledge bases or ontologies. User studies are obviously the most demanding in terms of efforts both on the data analysts and the users. It has thus been argued that insights from social and cognitive sciences (Miller 2019; Kliegr et al. 2021) should receive higher attention in the design of suitable criteria for assessing the interpretability of models and explanations.

3.2 Explaining black-box versus learning white-box models

The excellent performance of black-box models in many domains has resulted in the demand for XAI algorithms that aim at explaining such models, either via methods that are targeted to specific types of black-box models such as deep neural networks (Samek et al. 2019), or general methods that can take any black-box model such as LIME (Ribeiro et al. 2016) and (Lundberg and Lee 2017b). However, this approach has also been criticized for various reasons, including a lack of robustness (Slack et al. 2020; Zhang et al. 2019). In particular, Rudin (2019) has argued that surrogate explanatory models often do not faithfully capture the original black-box models, or are no more accurate than white-box models that have been directly learned from the original data, so that it seems advisable to invest more efforts into learning explainable white-box models in the first place. This comes back to learning simple (linear) scoring models (e.g., Puppe 2000; Atzmueller et al. 2006), which are interpretable and used, for example, in medicine as *diagnostic scores* (Ohmann et al. 1999; Fronhöfer and Schramm 2001). Here, Ustun and Rudin (2016, 2019) provided advanced algorithms for learning such diagnostic scores with optimized models that are still easily interpretable. Consequently, also based on such recent advances, research on the direct learning of white-box models has received a boost. For example, inductive rule learning has also seen a considerable increase of attention in recent years (see, e.g., Wang et al. 2017; Lakkaraju et al. 2016). New algorithms can also be tailored

to the specific challenges of providing an explanation. For example, the recent very efficient rule learner LORD (Huynh et al. 2023) draws inspiration from LIME and its rule-based version LORE (Guidotti et al. 2019) in that it learns the best possible rule for each training example in a similar way as XAI methods produce explanations that are tailored to particular query examples.

3.3 Explainability/interpretability on complex data

Interpretability and explainability are also an important emerging topic on complex data representations like network/graph and sequential data—with few works reviewing this field, yet. Recent work (e.g., Theissler et al. 2022; Zhao et al. 2023) focuses on the implementation of explainability on complex data, such as time series and networks/graphs. Here, prominent approaches also include post-hoc methods on graph neural networks (Liu et al. 2022; Huang et al. 2022; Schwenke et al. 2023). Regarding the modeling and explainable analysis of graphs and networks, for example, Masiala and Atzmueller (2018) outline perspectives on explanation in complex network analysis. As one of the emerging directions, we can consider the area of *anomaly detection* (Chandola et al. 2009). Here, the integration of feature-rich networks in (Interdonato et al. 2019) as well as methods for model-based anomaly link pattern mining have been presented, in particular also connected to according explanation methods (Guven et al. 2021). For example, we can apply *interpretable pattern mining* on graph structures, where the graph relations can be presented as explanations, e.g., for specific events and/or behavior in cyber security (Atzmueller et al. 2023).

Regarding time series and sequential data, examples are given by (Lonjarret et al. 2020; Iferroudjene et al. 2022) using local pattern mining; specifically for time series data (Vollert et al. 2021) methods include, for example, symbolic abstraction and/or transformer adaptations on time series data (Hsu et al. 2019; Schwenke and Atzmueller 2021b, a) for enhancing their interpretability, or signal processing techniques (Mochaourab et al. 2022).

3.4 Including background knowledge for explainability and interpretability

Background knowledge can potentially facilitate interpretability or explainability. If knowledge about relevant features or sub-concepts of a domain is available, this knowledge can be used to guide model construction and result in a more interpretable model. Alternatively, knowledge can be applied for providing (post-hoc) explanations, in a reconstructive manner both supporting as well as guiding the explanation generation process, such as enabling reconstructive explanations (Wick and Thompson 1992; Guven et al. 2021).

While at the beginning of deep learning research, the focus was exclusively on architectures for data-intensive neural network approaches, over the last years, there is a growing recognition that combining methods of knowledge-based artificial intelligence and machine learning is profitable for both areas (d'Avila Garcez

et al. 2019): On the one hand, machine learning can support data-based adaptation of predefined semantic models, on the other hand, knowledge can be used to guide and constrain model induction. Combining approaches of knowledge representation and reasoning with machine learning is often called hybrid AI or neuro-symbolic AI (Hitzler and Sarker 2022; d'Avila Garcez and Lamb 2023; De Raedt et al. 2019), e.g., using knowledge provided via knowledge graphs (Tiddi and Schlobach 2022).

Neural-symbolic AI aims to integrate two fundamental cognitive abilities—learning from experience and reasoning from what has been learned (d'Avila Garcez et al. 2008). That is, neuro-symbolic approaches provide computational architectures to combine system 1 and 2 thinking (Kahneman 2011).

An approach which naturally combines reasoning and learning is inductive logic programming (ILP) which was introduced in the early 1990s (Muggleton and De Raedt 1994; Cropper and Dumančić 2022). ILP is a family of approaches for learning relational models in the form of Prolog programs. Training examples are presented together with background knowledge given as ground facts. Additionally, background theories can be used. For instance, temporal or spatial calculi can be provided to include temporal or spatial relations in model induction (Katzouris et al. 2015; Bruckert et al. 2020). ILP is a special variant of inductive programming—that is, learning computer programs from examples (Gulwani et al. 2015). Learned models which are available in the form of computer programs are inherently interpretable since they are represented in symbolic form.

It has been shown that ILP models can help humans to comprehend complex relational dependencies (Muggleton et al. 2018). This characteristic has been introduced as ultra-strong machine learning by Michie (1988). Although ILP models are interpretable, to support human decision making in complex domains, it nevertheless is necessary to provide explanations which fulfill a current need for information. A local, verbal explanation can be extracted from the reasoning trace of the model for a given input (Gromowski et al. 2020; Finzel et al. 2021).

Different approaches have been proposed to combine symbolic approaches to reasoning and learning with deep learning approaches: One type uses deep learned models like sensors for perceptual information and constructs interpretable models on top. Another type of models incorporates knowledge into neural network architectures. Two examples for the first type are (a) the combination of CNNs with ILP conveyed via the model agnostic explanatory system LIME (Rabold et al. 2019) and (b) the combination of deep learning and probabilistic logic programming, for instance to learn arithmetic operations from hand written numbers (Manhaeve et al. 2018).

Different propositions to encode knowledge in neural networks are graph neural networks (Wu et al. 2020), differentiable inductive logic (Evans and Grefenstette 2018), and logic tensor networks (Badreddine et al. 2022).

4 Articles in this special issue

"Opening the black box" is an increasingly critical step in machine learning workflows. It helps to validate predictive models and retain user trust in AI. It also has utility in data mining, where it can help to extract interpretable insights from black

box models. However, a key aspect is the reliability of the explanations. A path towards higher quality explanations involves integrating knowledge from various sources and being able to process the knowledge in a logically sound way. We, therefore, invited submissions covering the use of knowledge graphs in XAI research as well as reasoning, the causality of machine learning models and logic programming.

As a response to the call for papers to this special issue, first issued in the fall of 2020, we received a total of 81 submissions. Of these, 24 papers were accepted, 5 of which have already been published in earlier issues of this journal (Volume 36 3/4), so that this printed issue contains a total of 19 articles. However, all articles are collected in a **topical collection** of this journal,[3] which as of now also contains an additional survey article, and may be further extended in the future.

In the following, we give a brief overview of the 25 articles, which are as of now contained in the on-line topical collection, loosely grouped according to the type of research problems addressed in the papers.

Surveys

A large number of surveys of research in XAI have been published in recent years. Nevertheless we have included three surveys in this topical collection, which we believe provide a novel and useful angle to the field. Concerning that, Schwalbe and Finzel (2023) provide a meta-survey, i.e., they evaluated and compared various surveys with the goal of providing a unified view, as well as a complete taxonomy of XAI methods. Guidotti (2023) provides an overview of work on counterfactual explanations, a popular family of methods which focus on identifying examples that are similar to a query example, but belong to a different class. This may also serve as an introduction to several papers in this special issue which also touch upon this topic (see below). Finally, we also included a paper (Bodria et al. 2023) in this topical collection, which provides a survey on the well investigated field of explaining black-box models, but provides a new angle by categorizing the work according to the type of explanation produced.

Feature-based Explanations

Arguably the best-known type of XAI methods are feature-based, such as LIME and SHAP. Typically, such techniques provide weights for features, which indicate their respective importance for the decision outcome. Not surprisingly, a fair share of works in this topical collection is devoted to this type of problem. Vreš and Robnik-Šikonja (2023) focus on a known problem with such techniques, namely that they could be manipulated via the choice of the employed feature sampling method. As a remedy, they propose a technique for modeling the data distribution and sampling according to the found distribution. Molnar et al. (2023) also notice a problem with one particular XAI technique, namely feature permutations. In particular, they observe that such permutations may sometimes generate data points in vacuous and meaningless parts of the instance space, and address that problem by first identifying feature groups in which no dependencies exist. The assessment of the

[3] https://link.springer.com/journal/10618/topicalCollection/AC_81d874a4efbe7c4f2ebf44b6cccb49df.

importance of feature groups as opposed to individual features is central to the work of Au et al. (2022), who survey possible ways for adapting various feature-based XAI techniques to that problem, and also introduce an appropriate visualization for feature groups. Somewhat orthogonally, Brandsæter and Glad (2023) generalize Shapley values, which are typically used for quantifying the contribution of individual features to the prediction of a single example, to a setting where they can be used to explain clusters or groups of examples. Scholbeck et al. (2024) argue that while the coefficients of linear regression models, which are, e.g., used by LIME, can be interpreted as feature importance weights, this does not hold for non-linear generalized regression models. As a remedy, the authors adapt a known statistical technique—marginal effects—to the problem of providing post-hoc explanations for predictions, and show that this compares favorably to standard linear techniques.

Counterfactual Explanations

While feature-based methods focus on individual features or feature groups, exemplar-based methods focus on entire examples. The most prominent approach in that class are counterfactual explanations, which aim at identifying an example of a different class, similar to the one that should be explained. The remaining differences between these two examples may be interpreted as a set of features which are most responsible for the provided decision, because changing them would change the outcome. As already mentioned, Guidotti (2023) provides an exhaustive overview and comparison of such methods, which is a good starting point for the following research contributions. Brughmans et al. (2023) introduce a novel technique for finding counterfactual examples, which has several attractive properties, including that it can provide multiple explanations that optimize different criteria. The approach is shown to outperform its competitors in a broad empirical study on 40 datasets and 3 assessment criteria. Crupi et al. (2023) criticize that, in general, work in this area focuses only on features, ignoring the causal impact of actions that need to be taken in order to change the outcome, and propose a general methodology that can be used for taking causal models into account. Raimundo et al. (2023) search for an optimal subset of counterfactual features for a desired outcome in a process they call counterfactual antecedent mining, which shares some similarities with rule mining algorithms, and show that it can produce Pareto-optimal solutions in the presence of multiple objectives. Guidotti et al. (2023) suggest the use of actionable counterfactual rules, which may be viewed as a generalization of counterfactual examples, to complement a conventional explanation in the form of a logic rule. The technique may thus be viewed as a hybrid bridge to the model-based explanation techniques covered below.

Model-based Explanations

Zhou et al. (2023) focus on model distillation, i.e., the process of training an interpretable surrogate model in lieu of a non-interpretable black-box model. In particular, they show that extracted decision-tree models can vary substantially, and propose a framework with which more stable models can be learned. While this work learns global symbolic models, Mollas et al. (2022) propose a technique for extracting local, rule-based explanations, in their case specifically from ensembles

of decision trees, also known as random forests. Veyrin-Forrer et al. (2023) aim at characterizing the behavior of graph neural networks via rule-based activation patterns. To find these, techniques from subgroup discovery are adapted to an XAI framework. Finally, Javed et al. (2022) focus on a class of methods that has so far not received much attention in XAI, namely the learning of computational models via genetic programming. Such methods are also susceptible to learning overly specific models, which can be countered by a variety of techniques, which are discussed in this introductory survey.

XAI for Deep Learning

Deep neural networks (DNNs) are as of now the most popular and arguably the most powerful black-box machine learning technique. Not surprisingly, a variety of XAI techniques have been developed specifically for understanding such models. Hada et al. (2023) propose to explain the last layers of a neural network with an interpretable decision tree model, with the goal of obtaining insight into what parts of the DNN are responsible for the observed prediction. This focus on the internal features of the network is, in a way, orthogonal to conventional XAI techniques, which focus on the features in the data. Ventura et al. (2023) focus on convolutional neural networks and propose an innovative technique for providing visual explanations based on unsupervised feature extraction from multiple convolutional layers. Schneider and Vlachos (2023) propose an interesting technique how explanations can be used for improved classification performance, in a process that is motivated by human reflective reasoning. The idea is to compute an explanation for an example for any class in a backwards relevance propagation up to a certain layer, and use this information as an additional input to a second deep network which is then able to make a prediction based on the input and the inferred explanation. Merz et al. (2022) propose a novel technique for obtaining feature-based explanations, which is, in principle, model-agnostic, but assumes that gradients can be computed, which makes it particularly well suited for explaining DNNs. The key innovation is to base the estimation on the quantiles of the prediction function, which also allows to obtain explanations for different levels of the response.

User-centric XAI

Many state-of-the-art techniques ignore the user and provide the same explanation irrespective of the target audience. Several papers address this issue by including the user into the explanation process. Sovrano and Vitali (2023) introduce a software library that aims at improving the usability of a wide range of XAI systems. The work takes a user-centred explanatory approach, founded in a sound theoretical basis (Achinstein's theory of explanations), and demonstrates how their work goes beyond user-agnostic textual or visual explanations. A similar point of view is also taken by Sokol and Flach (2024), who analyze the potential of various interpretable representations for adapting the presented information to particular users or applications. Baniecki et al. (2023) aim at explicitly modelling the interactive process by which an XAI-supported user may gain insight into a model's behavior and demonstrate the utility of this methodology for human decision making in a user study.

Applications of XAI

Finally, we were happy to receive a selection of studies on XAI techniques in real-world problems. Coma-Puig et al. (2023) report on the incorporation of standard XAI techniques such as LIME and SHAP into a system for the detection of so-called non-technical losses in electricity and gas transmission, i.e., losses that are not due to technical failures but to other factors such as fraudulent customers or malfunctioning meters. Valente et al. (2022) is somewhat orthogonal to the above approach, in that it reports on the design of a novel XAI technique, specifically targeted towards a clinical decision support system. The key idea is to first learn a set of interpretable decision rules, and then a neural network model which can be used for personalizing the rules for a given patient.

5 Conclusions: the future of explanations

The topic of explainability of AI systems has a long tradition and has been researched already in the context of expert systems (Clancey 1983). Here, mostly rule-based and verbal explanations have been researched. With the rise of deep learning, there is an ever growing set of XAI methods from feature relevance to contrastive explanations. The different approaches are so varied that multiple taxonomies for their categorization have been proposed (Arrieta et al. 2020; cf. also Schwalbe and Finzel (2023) in this issue). We believe that the breadth of articles in this topical collection is a representative snapshot of the state of the art in this area. Nevertheless, there is a need for further research in XAI (Schmid and Wrede 2022), and we will conclude with a (certainly subjective) outlook on the future of the field.

Of particular importance is to provide methods to evaluate the quality of generated explanations. On the one hand, the quality of explanations refers to the relation between the learned model and its explanation with faithfulness as most important evaluation criterion (Lakkaraju et al. 2019). On the other hand, quality of explanations refers to the human recipient of the explanation. The current practice of evaluating explanations by conveniently measurable quantities such as fidelity to a black-box model or complexity of the provided surrogate model does only superficially address the latter point. A core difficulty for interpretable models and ex-post explanations alike is that the quality of an explanation is context-dependent: It depends on what is explained to whom in what way (how) and for what reason (why). Research on tailoring explanations to the specific information needs of humans is still at its beginning and demands a broader, interdisciplinary perspective (Göbel et al. 2022; Kliegr et al. 2021). For example, a caveat with existing XAI approaches is that many generate explanations which are understandable only to experts and may require training to be used effectively, a problem that may also apply to white-box models (Reed et al. 2021).

The next evolutionary step, also reflecting what may be legally expected, are interactive explanations in natural language that involve reasoning. Such explanations may appear more convincing and informative as well as legally compliant.

However, with the current state of technology, the quality of such explanations may outpace the reliability of the predictions that are being explained. To illustrate this, the GPT-3 Chatbot, when asked why a kilo of beef weighs more than a kilo of air, provided the following explanation (Hern 2022):

> The weight of an object is determined by its mass, and the mass of a substance is a measure of the amount of matter it contains. Beef contains more matter than compressed air, so it has a higher mass and therefore weighs more.

This explanation refers to concepts from physics, is written in a precise, grammatical language, and as such, may appear more convincing than, for example, feature scores generated by SHAP and LIME, yet it is wrong. This example seems to support a pattern described by Saha et al. (2022), who show that GPT-3 Chatbot explanations are as "grammatical" as human explanations but significantly worse in generalizability and support of the label. Nevertheless, ChatGPT provides the opportunity to interactively refine and tune a provided explanation in a very natural way, as most of our readers have certainly experienced. We believe that such a seamless approach for interactive refinement of explanations is of utmost importance and will become a dominant trend in the field.

Returning to the quotation by Asimov, with which we motivated the need for explanations in Sect. 2.1: in the same story, when a robot was asked for a particular justification, it replies *"The matter admits no explanation."* As further elaborated by Asimov, such a reply would be valid if there was not enough data for a definite answer. Accompanying a wrong prediction with a convincing explanation (as in the ChatGPT example above) may be worse than no explanation, a phenomenon already Asimov was aware of. A future of XAI might be providing Asimov's "narrative" explanations in natural language, or maybe no explanation at all?

Funding Open Access funding enabled and organized by Projekt DEAL.

Papers in this Topical Collection

Au Q, Herbinger J, Stachl C et al (2022) Grouped feature importance and combined features effect plot. Data Min Knowl Disc 36(4):1401–1450. https://doi.org/10.1007/s10618-022-00840-5

Baniecki H, Parzych D, Biecek P (2023) The grammar of interactive explanatory model analysis. Data Min Knowl Disc. https://doi.org/10.1007/s10618-023-00924-w

Bodria F, Giannotti F, Guidotti R et al (2023) Benchmarking and survey of explanation methods for black box models. Data Min Knowl Disc 37(5):1719–1778. https://doi.org/10.1007/s10618-023-00933-9

Brandsæter A, Glad IK (2023) Shapley values for cluster importance. Data Min Knowl Disc. https://doi.org/10.1007/s10618-022-00896-3
Brughmans D, Leyman P, Martens D (2023) NICE: an algorithm for nearest instance counterfactual explanations. Data Min Knowl Disc. https://doi.org/10.1007/s10618-023-00930-y
Coma-Puig B, Calvo A, Carmona J et al (2023) A case study of improving a non-technical losses detection system through explainability. Data Min Knowl Disc. https://doi.org/10.1007/s10618-023-00927-7
Crupi R, Castelnovo A, Regoli D et al (2023) Counterfactual explanations as interventions in latent space. Data Min Knowl Disc. https://doi.org/10.1007/s10618-022-00889-2
Guidotti R (2023) Counterfactual explanations and how to find them: literature review and benchmarking. Data Min Knowl Disc. https://doi.org/10.1007/s10618-022-00831-6
Guidotti R, Monreale A, Ruggieri S et al (2023) Stable and actionable explanations of black-box models through factual and counterfactual rules. Data Min Knowl Disc. https://doi.org/10.1007/s10618-022-00878-5
Hada SS, Carreira-Perpiñán MÁ, Zharmagambetov A (2023) Sparse oblique decision trees: a tool to understand and manipulate neural net features. Data Min Knowl Disc. https://doi.org/10.1007/s10618-022-00892-7
Javed N, Gobet F, Lane P (2022) Simplification of genetic programs: a literature survey. Data Min Knowl Disc 36(4):1279–1300. https://doi.org/10.1007/s10618-022-00830-7
Merz M, Richman R, Tsanakas A et al (2022) Interpreting deep learning models with marginal attribution by conditioning on quantiles. Data Min Knowl Disc 36(4):1335–1370. https://doi.org/10.1007/s10618-022-00841-4
Mollas I, Bassiliades N, Tsoumakas G (2022) Conclusive local interpretation rules for random forests. Data Min Knowl Disc 36(4):1521–1574. https://doi.org/10.1007/s10618-022-00839-y
Molnar C, König G, Bischl B et al (2023) Model-agnostic feature importance and effects with dependent features: a conditional subgroup approach. Data Min Knowl Disc. https://doi.org/10.1007/s10618-022-00901-9
Raimundo MM, Nonato LG, Poco J (2023) Mining Pareto-optimal counterfactual antecedents with a branch-and-bound model-agnostic algorithm. Data Min Knowl Disc. https://doi.org/10.1007/s10618-022-00906-4
Schneider J, Vlachos M (2023) Reflective-net: learning from explanations. Data Min Knowl Disc. https://doi.org/10.1007/s10618-023-00920-0
Scholbeck CA, Casalicchio G, Molnar C et al (2024) Marginal effects for non-linear prediction functions. Data Min Knowl Disc. https://doi.org/10.1007/s10618-023-00993-x
Schwalbe G, Finzel B (2023) A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. Data Min Knowl Disc. https://doi.org/10.1007/s10618-022-00867-8
Sokol K, Flach P (2024) Interpretable representations in explainable AI: from theory to practice. Data Min Knowl Disc. https://doi.org/10.1007/s10618-024-01010-5
Sovrano F, Vitali F (2023) ExplanatorY Artificial Intelligence (YAI): human-centered explanations of explainable AI and complex data. Data Min Knowl Disc. https://doi.org/10.1007/s10618-022-00872-x
Valente F, Paredes S, Henriques J et al (2022) Interpretability, personalization and reliability of a machine learning based clinical decision support system. Data Min Knowl Disc 36(3):1140–1173. https://doi.org/10.1007/s10618-022-00821-8
Ventura F, Greco S, Apiletti D et al (2023) Explaining deep convolutional models by measuring the influence of interpretable features in image classification. Data Min Knowl Disc. https://doi.org/10.1007/s10618-023-00915-x
Veyrin-Forrer L, Kamal A, Duffner S et al (2023) On GNN explainability with activation rules. Data Min Knowl Disc. https://doi.org/10.1007/s10618-022-00870-z
Vreš D, Robnik-Šikonja M (2023) Preventing deception with explanation methods using focused sampling. Data Min Knowl Disc. https://doi.org/10.1007/s10618-022-00900-w
Zhou Y, Zhou Z, Hooker G (2023) Approximation trees: statistical reproducibility in model distillation. Data Min Knowl Disc. https://doi.org/10.1007/s10618-022-00907-3

References

Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6:52,138-52,160

Ahmed I, Jeon G, Piccialli F (2022) From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where. IEEE Trans Ind Inform 18(8):5031–5042

Ai L, Muggleton SH, Hocquette C et al (2021) Beneficial and harmful explanatory machine learning. Mach Learn 110:695–721

Ali S, Abuhmed T, El-Sappagh S et al (2023) Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. Inf Fusion 99(101):805

Allahyari H, Lavesson N (2011) User-oriented assessment of classification model understandability. In: Kofod-Petersen A, Heintz F, Langseth H (eds) Proceedings of the 11th Scandinavian conference on artificial intelligence (SCAI-11). IOS Press, Trondheim, Norway, pp 11–19

Arrieta AB, Díaz-Rodríguez N, Del Ser J et al (2020) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58:82–115

Asimov I (1942) Runaround. Astounding Sci Fict 29(1):94–103

Asimov I (1950) The evitable conflict. Astounding Sci Fict 45(4):48–68

Atzmueller M, Puppe F, Buscher HP (2005) Profiling examiners using intelligent subgroup mining. In: Proceedings of the 10th international workshop on intelligent data analysis in medicine and pharmacology (IDAMAP-2005), Aberdeen, Scotland, pp 46–51

Atzmueller M, Baumeister J, Puppe F (2006) Semi-automatic learning of simple diagnostic scores utilizing complexity measures. Artif Intell Med 37(1):19–30

Atzmueller M, Sylvester S, Kanawati R (2023) Exploratory and explanation-aware network intrusion profiling using subgroup discovery and complex network analysis. In: Proceedings of the European interdisciplinary cybersecurity conference (EICC). ACM, pp 153–158

Bach S, Binder A, Montavon G et al (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE 10(7):e0130140

Badreddine S, d'Avila Garcez AS, Serafini L et al (2022) Logic tensor networks. Artif Intell 303(103):649

Bansal G, Wu T, Zhou J et al (2021) Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In: Proceedings of the 2021 CHI conference on human factors in computing systems, pp 1–16

Bauer K, von Zahn M, Hinz O (2023) Expl(AI)ned: the impact of explainable artificial intelligence on cognitive processes. Information systems research

Bogert E, Schecter A, Watson RT (2021) Humans rely more on algorithms than social influence as a task becomes more difficult. Sci Rep 11(1):1–9

Bruckert S, Finzel B, Schmid U (2020) The next generation of medical decision support: a roadmap toward transparent expert companions. Front Artif Intell 3(507):973

Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: a survey. ACM Comput Surv 41(3):1–58

Cian D, van Gemert J, Lengyel A (2020) Evaluating the performance of the LIME and Grad-CAM explanation methods on a lego multi-label image classification task. arXiv preprint arXiv:2008.01584

Clancey WJ (1983) The epistemology of a rule-based expert system: a framework for explanation. Artif Intell 20(3):215–251

Cohen WW (1995) Fast effective rule induction. In: Prieditis A, Russell S (eds) Proceedings of the 12th international conference on machine learning (ML-95). Morgan Kaufmann, Lake Tahoe, CA, pp 115–123

Craven MW, Shavlik JW (1995) Extracting tree-structured representations of trained networks. In: Touretzky DS, Mozer M, Hasselmo ME (eds) Advances in neural information processing systems 8 (NIPS 1995). MIT Press, Cambridge, pp 24–30

Cropper A, Dumančić S (2022) Inductive logic programming at 30: a new introduction. J Artif Intell Res 74:765–850

Cropper A, Dumančić S, Evans R et al (2022) Inductive logic programming at 30. Mach Learn 111(1):147–172

d'Avila Garcez A, Lamb LC (2023) Neurosymbolic AI: the 3rd wave. Artif Intell Rev 56(11):12,387-12,406

d'Avila Garcez AS, Lamb LC, Gabbay DM (2008) Neural-symbolic cognitive reasoning. Springer, Berlin

d'Avila Garcez AS, Gori M, Lamb LC et al (2019) Neural-symbolic computing: an effective methodology for principled integration of machine learning and reasoning. J Appl Log J 6(4):611–632
De Raedt L, Manhaeve R, Dumančić S et al (2019) Neuro-symbolic = neural + logical + probabilistic. In: The 14th international workshop on neural-symbolic learning and reasoning (NeSy-19@IJCAI)
De Raedt L, Dumančić S, Manhaeve R et al (2020) From statistical relational to neuro-symbolic artificial intelligence. In: Bessiere C (ed) Proceedings of the twenty-ninth international joint conference on artificial intelligence, IJCAI 2020. ijcai.org, pp 4943–4950
Deeks A (2019) The judicial demand for explainable artificial intelligence. Columbia Law Rev 119(7):1829–1850
Dhurandhar A, Chen P, Luss R et al (2018) Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In: Bengio S, Wallach HM, Larochelle H et al (eds) Advances in neural information processing systems 31 (NeurIPS 2018), pp 590–601
Evans R, Grefenstette E (2018) Learning explanatory rules from noisy data. J Artif Intell Res 61:1–64
Fayyad UM, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery: an overview. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P et al (eds) Advances in knowledge discovery and data mining. AAAI Press, pp 1–34
Fernández-Delgado M, Cernadas E, Barro S et al (2014) Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res 15(1):3133–3181
Finzel B, Tafler DE, Scheele S et al (2021) Explanation as a process: user-centric construction of multi-level and multi-modal explanations. In: Proceedings of the 44th German conference on artificial intelligence (KI). Springer, pp 80–94
Freitas AA (2019) Automated machine learning for studying the trade-off between predictive accuracy and interpretability. In: Proceedings of the international cross-domain conference for machine learning and knowledge extraction (CD-MAKE). Springer, pp 48–66
Fronhöfer B, Schramm M (2001) A probability theoretic analysis of score systems. In: Kern-Isberner G, Lukasiewicz T, Weydert E (eds) KI-2001 workshop: uncertainty in artificial intelligence, pp 95–108
Fürnkranz J, Knobbe A (2010) Guest editorial: global modeling using local patterns. Data Min Knowl Disc 21:1–8
Fürnkranz J, Gamberger D, Lavrač N (2012) Foundations of rule learning. Springer, Berlin
Fürnkranz J, Kliegr T, Paulheim H (2020) On cognitive preferences and the plausibility of rule-based models. Mach Learn 109(4):853–898
Gabriel A, Paulheim H, Janssen F (2014) Learning semantically coherent rules. In: Cellier P, Charnois T, Hotho A et al (eds) Proceedings of the ECML/PKDD-14 international workshop on interactions between data mining and natural language processing. In: CEUR workshop proceedings, Nancy, France, pp 49–63
Gade K, Geyik SC, Kenthapadi K et al (2019) Explainable AI in industry. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining, pp 3203–3204
Göbel K, Niessen C, Seufert S et al (2022) Explanatory machine learning for justified trust in human-AI collaboration: experiments on file deletion recommendations. Front Artif Intell 5(919):534
Górriz J, Álvarez-Illán I, Álvarez-Marquina A et al (2023) Computational approaches to explainable artificial intelligence: advances in theory, applications and trends. Inf Fusion 100(101):945
Gromowski M, Siebers M, Schmid U (2020) A process framework for inducing and explaining datalog theories. Adv Data Anal Classif 14(4):821–835
Guidotti R, Monreale A, Giannotti F et al (2019) Factual and counterfactual explanations for black box decision making. IEEE Intell Syst 34(6):14–23
Gulwani S, Hernández-Orallo J, Kitzelmann E et al (2015) Inductive programming meets the real world. Commun ACM 58(11):90–99
Gunning D, Stefik M, Choi J et al (2019) XAI: explainable artificial intelligence. Sci Robot 4:eaay7120
Guven C, Seipel D, Atzmueller M (2021) Applying ASP for knowledge-based link prediction with explanation generation in feature rich networks. IEEE Trans Netw Sci Eng 8(2):1305–1315
Hedström A, Weber L, Krakowczyk D et al (2023) Quantus: an explainable AI toolkit for responsible evaluation of neural network explanations and beyond. J Mach Learn Res 24(34):1–11
Herchenbach M, Müller D, Scheele S et al (2022) Explaining image classifications with near misses, near hits and prototypes: supporting domain experts in understanding decision boundaries. In: International conference on pattern recognition and artificial intelligence. Springer, pp 419–430

Hern A (2022) Techscape: meet ChatGPT, the viral AI tool that may be a vision of our weird tech future. The Guardian https://www.theguardian.com/technology/2022/dec/06/meet-chatgpt-the-viral-ai-tool-that-may-be-a-vision-of-our-weird-tech-future

Hitzler P, Sarker M (eds) (2022) Neuro-symbolic artificial intelligence: the state of the art. IOS Press, Amsterdam

Holzinger A, Biemann C, Pattichis CS et al (2017) What do we need to build explainable AI systems for the medical domain? In: Proceedings of the biomedical NLP workshop associated with RANLP 2017, Varna, Bulgaria, pp 42–48. arXiv preprint arXiv:1712.09923

Hooker S, Erhan D, Kindermans PJ et al (2019) A benchmark for interpretability methods in deep neural networks. In: Advances in neural information processing systems, vol 32

Hsu EY, Liu CL, Tseng VS (2019) Multivariate time series early classification with interpretability using deep learning and attention mechanism. In: Pacific-Asia conference on knowledge discovery and data mining. Springer, pp 541–553

Huang X, Kroening D, Ruan W et al (2020) A survey of safety and trustworthiness of deep neural networks: verification, testing, adversarial attack and defence, and interpretability. Comput Sci Rev 37(100):270

Huang Q, Yamada M, Tian Y et al (2022) Graphlime: local interpretable model explanations for graph neural networks. IEEE Trans Knowl Data Eng 35:6968–6972

Huynh VQP, Fürnkranz J, Beck F (2023) Efficient learning of large sets of locally optimal classification rules. Mach Learn 112(2):571–610

Huysmans J, Dejaeger K, Mues C et al (2011) An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. Decis Support Syst 51(1):141–154

Iferroudjene M, Lonjarret C, Robardet C et al (2022) Methods for explaining top-N recommendations through subgroup discovery. Data Min Knowl Disc 37(2):833–872

Interdonato R, Atzmueller M, Gaito S et al (2019) Feature-rich networks: going beyond complex network topologies. Appl Netw Sci 4(1):4

Kahneman D (2011) Thinking, fast and slow. Macmillan, New Yrok

Katzouris N, Artikis A, Paliouras G (2015) Incremental learning of event definitions with inductive logic programming. Mach Learn 100(2):555–585

Kaur D, Uslu S, Durresi A et al (2021) Trustworthy explainability acceptance: a new metric to measure the trustworthiness of interpretable AI medical diagnostic systems. In: Complex, intelligent and software intensive systems: proceedings of the 15th international conference on complex, intelligent and software intensive systems (CISIS-2021). Springer, pp 35–46

Kim B, Koyejo O, Khanna R (2016) Examples are not enough, learn to criticize! Criticism for interpretability. In: Lee DD, Sugiyama M, von Luxburg U et al (eds) Advances in neural information processing systems 29 (NeurIPS 2016), pp 2280–2288

Kliegr T, Bahník Š, Fürnkranz J (2021) A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. Artif Intell 295(103):458

Kononenko I (1993) Inductive and Bayesian learning in medical diagnosis. Appl Artif Intell 7:317–337

Lakkaraju H, Bach SH, Leskovec J (2016) Interpretable decision sets: a joint framework for description and prediction. In: Proceedings of the 22Nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, NY, USA, KDD '16, pp 1675–1684

Lakkaraju H, Kamar E, Caruana R et al (2019) Faithful and customizable explanations of black box models. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society, pp 131–138

Laux J, Wachter S, Mittelstadt B (2023) Trustworthy artificial intelligence and the European Union AI act: on the conflation of trustworthiness and acceptability of risk. Regul Gov 18:3–32

Lim S (2021) Judicial decision-making and explainable artificial intelligence: a reckoning from first principles. Singap Acad of Law J 33:280–314

Liu N, Feng Q, Hu X (2022) Interpretability in graph neural networks. In: Foundations, frontiers, and applications, graph neural networks, pp 121–147

Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: people prefer algorithmic to human judgment. Organ Behav Hum Decis Process 151:90–103

Lonjarret C, Robardet C, Plantevit M et al (2020) Why should I trust this item? Explaining the recommendations of any model. In: Proc. IEEE international conference on data science and advanced analytics (DSAA). IEEE, pp 526–535

Lundberg SM, Lee S (2017a) A unified approach to interpreting model predictions. In: Guyon I, von Luxburg U, Bengio S et al (eds) Advances in neural information processing systems, vol 30. Long Beach, CA, USA, pp 4765–4774

Lundberg SM, Lee S (2017b) A unified approach to interpreting model predictions. In: Guyon I, von Luxburg U, Bengio S et al (eds) Advances in neural information processing systems 30: annual conference on neural information processing systems (NeuRIPS 2017), pp 4765–4774
Manhaeve R, Dumančić S, Kimmig A et al (2018) DeepProbLog: neural probabilistic logic programming. In: Advances in neural information processing systems, vol 31
Masiala S, Atzmueller M (2018) First perspectives on explanation in complex network analysis. In: Proceedings of the Benelux conference on artificial intelligence (BNAIC), Jheronimus Academy of Data Science,'s-Hertogenbosch, The Netherlands
Michie D (1988) Machine learning in the next five years. In: Proceedings of the 3rd European working session on learning (EWSL), pp 107–122
Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1–38
Mochaourab R, Venkitaraman A, Samsten I et al (2022) Post hoc explainability for time series classification: toward a signal processing perspective. IEEE Signal Process Mag 39(4):119–129
Molnar C (2022) Interpretable machine learning: a guide for making black box models explainable, 2nd edn. http://christophm.github.io/interpretable-ml-book/
Morik K, Boulicaut JF, Siebes A (eds) (2005) Local pattern detection. Springer, Berlin
Muggleton S, De Raedt L (1994) Inductive logic programming: theory and methods. J Log Program 19:629–679
Muggleton SH, Schmid U, Zeller C et al (2018) Ultra-strong machine learning: comprehensibility of programs learned with ILP. Mach Learn 107:1119–1140
Nadeem A, Vos D, Cao C et al (2023) Sok: explainable machine learning for computer security applications. In: Proceedings is the IEEE 8th European symposium on security and privacy (EuroS &P). IEEE, pp 221–240
Ohmann C, Franke C, Yang Q (1999) Clinical benefit of a diagnostic score for appendicitis: results of a prospective interventional study. Arch Surg 134:993–996
Palczewska A, Palczewski J, Robinson RM et al (2013) Interpreting random forest models using a feature contribution method. In: 2013 IEEE 14th international conference on information reuse & integration (IRI). IEEE, pp 112–119
Pazzani MJ, Mani S, Shankle WR (2001) Acceptance of rules generated by machine learning among medical experts. Methods Inf Med 40(5):380–385
Petsiuk V, Das A, Saenko K (2018) RISE: randomized input sampling for explanation of black-box models. In: British machine vision conference 2018, (BMVC 2018). BMVA Press, p 151
Piltaver R, Luštrek M, Gams M et al (2016) What makes classification trees comprehensible? Expert Syst Appl 62:333–346
Puppe F (2000) Knowledge formalization patterns. In: Proceedings of the PKAW 2000, Sydney, Australia
Rabold J, Deininger H, Siebers M et al (2019) Enriching visual with verbal explanations for relational concepts: combining LIME with Aleph. In: Proceedings of the European conference on machine learning and knowledge discovery in databases (ECML/PKDD). Springer, pp 180–192
Rabold J, Schwalbe G, Schmid U (2020) Expressive explanations of DNNs by combining concept analysis with ILP. In: Advances in artificial intelligence: proceedings of the 43rd German conference on AI (KI). Springer, pp 148–162
Rabold J, Siebers M, Schmid U (2022) Generating contrastive explanations for inductive logic programming based on a near miss approach. Mach Learn 111(5):1799–1820
Reed C, Grieman K, Early J (2021) Non-Asimov explanations regulating AI through transparency. arXiv preprint arXiv:2111.13041
Ribeiro MT, Singh S, Guestrin C (2016) "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, pp 1135–1144
Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1:206–215
Rudin C, Radin J (2019) Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. Harv Data Sci Rev 1(2):10–1162
Ryo M (2022) Explainable artificial intelligence and interpretable machine learning for agricultural data analysis. Artif Intell Agric 6:257–265
Saha S, Hase P, Rajani N et al (2022) Are hard examples also harder to explain? A study with human and model-generated explanations. In: Goldberg Y, Kozareva Z, Zhang Y (eds) Proceedings of the

2022 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp 2121–2131
Samek W, Montavon G, Vedaldi A et al (eds) (2019) Explainable AI: interpreting, explaining and visualizing deep learning. In: Lecture notes in computer science, vol 11700. Springer
Schallner L, Rabold J, Scholz O et al (2020) Effect of superpixel aggregation on explanations in LIME: a case study with biological data. In: Part I (ed) Machine Learning and knowledge discovery in databases: international workshops of ECML PKDD 2019, proceedings. Springer, pp 147–158
Schmid U (2023) Trustworthy artificial intelligence: comprehensible, transparent and correctable. In: Werthner H, Ghezzi C, Kramer J et al (eds) Introduction to digital humanism. Springer, Berlin, pp 151–164
Schmid U, Wrede B (2022) What is missing in XAI so far? An interdisciplinary perspective. KI Künstliche Intell Spec Issue Explaina AI 36(3–4):303–315
Schmidt P, Biessmann F (2019) Quantifying interpretability and trust in machine learning systems. In: Proceedings of the AAAI-19 workshop on network interpretability for deep learning. arXiv preprint arXiv:1901.08558
Schmid U, Zeller C, Besold T et al (2017) How does predicate invention affect human comprehensibility? In: Cussens J, Russo A (eds) Proceedings of the 26th international conference on inductive logic programming (ILP-16). Springer, London, UK, pp 52–67
Schoenherr JR, Abbas R, Michael K et al (2023) Designing AI using a human-centered approach: explainability and accuracy toward trustworthiness. IEEE Trans Technol Soc 4(1):9–23
Schoenke J, Aschenbruck N, Interdonato R et al (2021) Gaia-AgStream: an explainable AI platform for mining complex data streams in agriculture. In: Proceedings of the international conference on smart and sustainable agriculture (SSA). Springer, Berlin/Heidelberg, Germany
Schwenke L, Atzmueller M (2021a) Constructing global coherence representations: identifying interpretability and coherences of transformer attention in time series data. In: Proceedings of the 8th IEEE international conference on data science and advanced analytics, DSAA 2021, Porto, Portugal, Oct 6–9, 2021. IEEE, pp 1–12
Schwenke L, Atzmueller M (2021b) Show me what you're looking for: visualizing abstracted transformer attention for enhancing their local interpretability on time series data. In: Proceedings of the 34th international Florida artificial intelligence research society conference (FLAIRS-2021), FLAIRS, North Miami Beach, FL, USA
Schwenke L, Bloemheuvel S, Atzmueller M (2023) Identifying informative nodes in attributed spatial sensor networks using attention for symbolic abstraction in a GNN-based modeling approach. In: Proceedings of the 36th international Florida artificial intelligence research society conference (FLAIRS-2023), FLAIRS, Clearwater Beach, FL, USA
Selvaraju RR, Cogswell M, Das A et al (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626
Slack D, Hilgard S, Jia E et al (2020) Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM conference on AI, Ethics, and society, pp 180–186
Theissler A, Spinnato F, Schlegel U et al (2022) Explainable AI for time series classification: a review, taxonomy and research directions. IEEE Access 10:100700–100724
Tiddi I, Schlobach S (2022) Knowledge graphs as tools for explainable machine learning: a survey. Artif Intell 302(103):627
Towell GG, Shavlik JW (1994) Knowledge-based artificial neural networks. Artif Intell 70(1–2):119–165
Ustun B, Rudin C (2016) Supersparse linear integer models for optimized medical scoring systems. Mach Learn 102:349–391
Ustun B, Rudin C (2019) Learning optimized risk scores. J Mach Learn Res 20(150):1–75
Vilone G, Longo L (2021) Notions of explainability and evaluation approaches for explainable artificial intelligence. Inf Fusion 76:89–106
Vollert S, Atzmueller M, Theissler A (2021) Interpretable machine learning: a brief survey from the predictive maintenance perspective. In: Proceedings of the IEEE international conference on emerging technologies and factory automation (ETFA 2021), IEEE
Wachter S, Mittelstadt B, Russell C (2017) Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv J Law Technol 31:841
Wang T, Rudin C, Doshi-Velez F et al (2017) A Bayesian framework for learning rule sets for interpretable classification. J Mach Learn Res 18:70:1-70:37

Weiss G, Goldberg Y, Yahav E (2018) Extracting automata from recurrent neural networks using queries and counterexamples. In: Proceedings of the international conference on machine learning (ICML), PMLR, pp 5247–5256

Wick MR, Thompson WB (1992) Reconstructive expert system explanation. Artif Intell 54(1–2):33–70

Wu Z, Pan S, Chen F et al (2020) A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst 32(1):4–24

Xiong H, Li X, Li X et al (2022) Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond. Knowl Inf Syst 64(12):3197–3234

Zhang Y, Song K, Sun Y et al (2019) "Why should you trust my explanation?" Understanding uncertainty in LIME explanations. In: Proceedings of the AI for social good ICML workshop, Long Beach, United States. arXiv preprint arXiv:1904.12991

Zhao Z, Shi Y, Wu S et al (2023) Interpretation of time-series deep models: a survey. arXiv preprint arXiv:2305.14582

Authors and Affiliations

Martin Atzmueller[1,5] · Johannes Fürnkranz[2] · Tomáš Kliegr[3] · Ute Schmid[4]

✉ Martin Atzmueller
martin.atzmueller@uni-osnabrueck.de

Johannes Fürnkranz
juffi@faw.jku.at

Tomáš Kliegr
tomas.kliegr@vse.cz

Ute Schmid
ute.schmid@uni-bamberg.de

[1] Semantic Information Systems Group, Osnabrück University, Wachsbleiche 27, 49090 Osnabrück, Germany

[2] Institute for Application-Oriented Knowledge Processing (FAW), Johannes-Kepler University, Altenberger Straße 69, 4040 Linz, Austria

[3] Department of Information and Knowledge Engineering, Prague University of Economics and Business, W. Churchill Sq. 1938/4, 610101 Prague, Czech Republic

[4] Cognitive Systems, University of Bamberg, An der Weberei 5, 96045 Bamberg, Germany

[5] German Research Center for Artificial Intelligence (DFKI), Hamburger Straße 24, 49084 Osnabrück, Germany