Received: 22 November 2019 | Revised: 18 September 2020 | Accepted: 21 September 2020

DOI: 10.1002/widm.1391

OVERVIEW

WIREs DATA MINING AND KNOWLEDGE DISCOVERY WILEY

A historical perspective of explainable Artificial Intelligence

Roberto Confalonieri[1] | **Ludovik Coba**[1] | **Benedikt Wagner**[2] | **Tarek R. Besold**[3]

[1]Faculty of Computer Science, Free University of Bozen-Bolzano, Bozen-Bolzano, Italy

[2]Research Centre for Machine Learning, City University, London, UK

[3]Neurocat GmbH, Berlin, Germany

Correspondence
Roberto Confalonieri, Faculty of Computer Science, Free University of Bozen-Bolzano, Dominikanerplatz 3, Bozen-Bolzano I-39100, Italy.
Email: roberto.confalonieri@unibz.it

Abstract

Explainability in Artificial Intelligence (AI) has been revived as a topic of active research by the need of conveying safety and trust to users in the "how" and "why" of automated decision-making in different applications such as autonomous driving, medical diagnosis, or banking and finance. While explainability in AI has recently received significant attention, the origins of this line of work go back several decades to when AI systems were mainly developed as (knowledge-based) expert systems. Since then, the definition, understanding, and implementation of explainability have been picked up in several lines of research work, namely, expert systems, machine learning, recommender systems, and in approaches to neural-symbolic learning and reasoning, mostly happening during different periods of AI history. In this article, we present a historical perspective of Explainable Artificial Intelligence. We discuss how explainability was mainly conceived in the past, how it is understood in the present and, how it might be understood in the future. We conclude the article by proposing criteria for explanations that we believe will play a crucial role in the development of human-understandable explainable systems.

This article is categorized under:

Fundamental Concepts of Data and Knowledge > Explainable AI

Technologies > Artificial Intelligence

KEYWORDS

explainable AI, explainable recommender systems, interpretable machine learning, neural-symbolic reasoning

1 | INTRODUCTION

As of 2020, explainability has been identified as a key factor for adoption of AI systems in a wide range of contexts (Doshi-Velez & Kim, 2017; Lipton, 2018; Ribeiro, Singh, & Guestrin, 2016a). Discussion accompanying the increasingly common deployment of intelligent systems in application domains such as autonomous vehicles and transportation, medical diagnosis, or insurance and financial services have shown that when decisions are taken or suggested by

automated systems, it is essential for practical, social, and—with increasing frequency—legal reasons that an explanation can be provided to users, developers, and regulators.

As a case in point, the European Union's General Data Protection Regulation (GDPR) stipulates a right to obtain "*meaningful information about the logic involved*"—commonly interpreted as a "right to an explanation"[1]—for consumers affected by an automatic decision (Parliament and Council of the European Union, 2016).[2]

The reasons for equipping intelligent systems with explanatory capabilities are not limited to issues of user rights and of technology acceptance, though. Explainability is also required by designers and developers to enhance system robustness and to enable diagnostics to prevent bias, unfairness, and discrimination, as well as to increase trust by all users in *why* and *how* decisions are made.

Being able to provide an explanation of why a certain decision was made, has thus become a desirable property of intelligent systems (Doran, Schulz, & Besold, 2017). Explanations should help users in understanding the model of the system, in order to maintain it, and to use it effectively; they should also assist the user when debugging the model to prevent and rectify incorrect conclusions. In addition, explanations can serve educational purposes and be helpful for people in discovering and understanding novel concepts in an application domain. Finally, explanations are related to users' trust and persuasion, they should convey a sense of actionability, and convince users that the system's decisions are the most convenient for them.

Notwithstanding, there is no clear agreement about what an explanation is, nor what a *good* explanation entails. Its manifestations have been studied across different incarnation of AI systems and disciplines. The first notions of explainability in Artificial Intelligence had subsided together with that in expert systems after the mid-1980s (Buchanan & Shortliffe, 1984; Wick & Thompson, 1992), and have been brought back into the focus by recent successes in machine learning technology (Guidotti et al., 2018), for both autonomous (Nunes & Jannach, 2017) and human-in-the-loop systems (Holzinger, 2016; Holzinger, Plass, et al., 2019), with applications in recommender systems (Tintarev & Masthof, 2015), and approaches of neural-symbolic learning and reasoning (Garcez et al., 2015).

In this article, we look at the literature of Explainable Artificial Intelligence (XAI) from a historical perspective of traditional approaches as well as approaches currently being developed. The relevant literature is vast, and this article does not aim to be a complete overview of the XAI literature. For each of the perspectives, the reader can find more comprehensive literature reviews in machine learning and Deep Learning (Arrieta et al., 2020; Fernandez, Herrera, Cordon, Jose del Jesus, & Marcelloni, 2019; Guidotti et al., 2018; Mueller, Hoffman, Clancey, Emrey, & Klein, 2019), recommender systems (Nunes & Jannach, 2017; Tintarev & Masthof, 2015), and Neural-Symbolic Approaches (Garcez et al., 2015). The aim of the article is rather to provide an overview and discuss how different notions of explainability (resp. format of explanations) have been conceived, and to provide several examples.

The main contributions of this article are:

- To provide an overview of XAI, and how it is understood in expert systems, machine learning, recommender systems, and neural-symbolic learning and reasoning approaches.
- To provide the reader with a wide range of references, (s)he can use to gain a deeper understanding in the topic of XAI.

The article is organized as follows. In Section 2, we give an overview of the different notions of explainability that are subsequently addressed from different perspectives throughout the article. Section 3 describes two notions of explanations prominently represented in the expert system literature, namely explanations as line of reasoning and as problem-solving activities. In Section 4, we present how the notion of explanation is commonly understood in machine learning, as well as a few examples of such explanations. Section 5 discusses how explanations are conceptualized in the context of recommender systems. Section 6 identifies the increasingly popular perspective of Neural-Symbolic Learning and Reasoning as promising approach to explainability in AI systems. Section 7 provides a critical discussion and comparison of the different notions of explainability mentioned throughout the article, and introduces general desiderata for explainability and a set of challenges for the development of human-understandable explainable AI systems. Section 8 concludes the article.

2 | WHAT IS A (GOOD) EXPLANATION?

Defining what an explanation is remains a still open research question. In particular, determining the criteria for a *good* explanation as of today is an active debate in various fields, including cognitive science, computer science, psychology,

and philosophy (Confalonieri et al., 2019; Guidotti et al., 2018; Hoffman, Mueller, Klein, & Litman, 2018; Lipton, 2018; Lombrozo, 2016; Miller, 2019).

Miller (2019) articulates the link between discussion in the social sciences and explainability in AI, providing an in-depth survey on research on explanations in philosophy, psychology, and cognitive science. Three major findings were highlighted. First, explanations are *counterfactual*, and humans tend to understand why a certain event happened instead of some other events. Second, explanations are *selective* and focus on one or two possible causes—instead of all possible causes—for a decision or recommendation; that is, explanations should not overwhelm the user with too much information. Third, explanations are a *social* conversation and interaction for the purpose of transferring knowledge, implying that the explainer must be able to leverage the mental model of the explainee while engaging in the explanation process. While according to Miller (2019) these three points are key properties when building useful explanations, the different notions of explainability prevalent in XAI only recently started to take them into account.

Psychology researchers have studied and defined properties of explanations that are human-oriented. For instance, Lombrozo (2016) suggested that one needs to differentiate between distinct possible goals for explainability, while highlighting why and how human explanatory cognition provides crucial constraints for the design of XAI systems. Hilton (1990) pointed out that explanations imply social interactions, and that for machine-generated explanations, it is essential to associate *semantic* information with an explanation (or its components) for effective knowledge transmission to human users. Kulesza et al. (2013) investigated the relationship between certain properties of generated explanations and the fidelity of users' mental models, finding that completeness ultimately appears to be more important than soundness, and that oversimplification is detrimental to users' trust in an explanation.

Work in computer science hitherto focused to the most part on the mechanistic aspects of how explanations are generated (Guidotti et al., 2018). This includes not only approaches in machine learning and recommender systems, but also in knowledge-based systems. The types of explanations these systems are able to create—and, consequently, their properties—mainly depend on the type of reasoning employed in the system, namely, symbolic, subsymbolic, or hybrid.[3]

Symbolic reasoning systems draw conclusions or explain why a certain hypothesis holds based on a knowledge base—usually encoded as a set of production or symbolic rules—and an inference mechanism, such as deduction, abduction, or analogical reasoning (Doyle, Tsymbal, & Cunningham, 2003; Lacave & Diez, 2004; Mitchell, Keller, & Kedar-Cabelli, 1986). Explanations in these systems consist of either descriptions coupled to the reasoning trace of the system, or descriptions more coupled to the story behind the decision-making process of the system (Buchanan & Shortliffe, 1984; Wick & Thompson, 1992). In either case, metrics and desirable properties for these explanations are, for instance, *accuracy*, *adaptability*, and *comprehensibility*. While these explanations are typically meant to be a precise reconstruction of the system behavior, they also should be adaptable to match different user profiles. Indeed, lay users might be more interested in a less accurate but more understandable explanation, whereas expert users might prefer more technical and precise explanation formats.

Subsymbolic (or connectionist) reasoning systems are, generally speaking, those that rely on machine learning models in which representations are in most cases distributed and processing occurs simultaneously in multiple parallel channels. Unfortunately, these properties frequently bring about a certain black-box nature of the corresponding models. As a consequence, explanations in these systems often take the form of interpretable models that approximate or try to mimic the behavior of the black-box (Andrews, Diederich, & Tickle, 1995; Guidotti et al., 2018). An interpretable model allows users to understand how decisions are made by means of local or global post-hoc explanations (Guidotti et al., 2018). Such interpretable models are typically evaluated using metrics such as *accuracy* and *fidelity*. These metrics measure to what extent an interpretable model is able to maintain competitive levels of accuracy with respect to the original black-box model, and to what extent the model is able to accurately imitate a black-box predictor respectively. Additional metrics targetting the notion of *causability* of explanations have recently been introduced by (Holzinger, Langs, et al., 2019; Holzinger, Carrington, & Müller, 2020). Causability refers to the extent to which an explanation achieves a certain level of causal understanding in a specified context of use and is measured in terms of effectiveness, efficiency, satisfaction related to causal understanding and its transparency for a user. As an additional class of explanation approaches predominantly for black-box models, methods providing explanations based on counterfactuals (i.e., hypothetical input examples that show how a different decision or prediction could have been obtained) recently also moved into the focus of active research (see e.g., Mothilal, Sharma, & Tan, 2020).

A particular category of sub-symbolic reasoning systems are recommender systems. There is no clear consensus in the recommender systems literature on what makes for a good explanation (Nilashi, Jannach, & bin Ibrahim, O., Esfahani, M. D.,, & Ahmadi, H., 2016; Nunes & Jannach, 2017; Tintarev & Masthof, 2015). In fact, an explanation on

the recommendation can have different goals, and impact decision-makers differently (Coba, Rook, et al., 2019). For example, a tailored explanation can persuade or help a user in finding an item more efficiently (Tintarev & Masthof, 2015). When implementing an explanation, a usual approach is to first determine its objective. For instance, stakeholders might be interested in delivering *persuasive* explanations, since they increase the probability of acceptance or purchase of a recommended item (Nunes & Jannach, 2017). *Trustworthiness* is another desired property of an explanation, since users tend to return to and reuse systems that they trust (L. Chen & Pu, 2005). Moreover, *efficient*, *effective*, and *satisfying* explanations help the users in deciding fast and making good decisions and increase the ease of use, respectively (Tintarev & Masthof, 2015). *Transparency* fosters the understandability for the user of the underlying logic of the advice-giving systems, and *scrutability* allows the user to tell that the system is wrong. These properties are often correlated. For instance, transparent explanations should also be comprehensible, and are known to convey trust. For a detailed discussion about the relationships between characteristics we refer the reader to (Balog & Radlinski, 2020).

Hybrid or neural-symbolic systems are those systems that combine symbolic and sub-symbolic reasoning (Garcez et al., 2015). The sub-symbolic system is able to build predictive models using connectionist machine learning and processing large amounts of data, while the symbolic system is equipped with a rich representation of domain knowledge and can be used for higher-level, structured reasoning. These symbolic elements are used by the system to explain the decisions made by the sub-symbolic components. Also here, *accuracy* and *fidelity* are, once more, important metrics to measure the performance of an interpretable model; whereas *consistency* and *comprehensibility* are desirable properties of the produced explanations from the explainee's point of view. The domain knowledge can serve as basis for common-sense reasoning, and supports knowledge abstraction, refinement, and injection (Confalonieri, Eppe, Schorlemmer, Kutz, & Pen˜aloza, R.,, & Plaza, E., 2018; Lehmann & Hitzler, 2010). As such, the system has not only the capability to create explanations for the sub-symbolic parts, but also to change the explanations' level of accuracy and technicality depending on the user profile. Furthermore, the system can refine the extracted knowledge, and inject it back to the sub-symbolic system to improve its performance (Garcez, Broda, & Gabbay, 2001).

3 | EXPLANATIONS IN EXPERT SYSTEMS

Expert or knowledge-based systems are software systems augmented by expert or domain knowledge. They are considered as one of the first instantiations of AI systems. They were developed to support humans in making decisions in several domains (Doyle et al., 2003; Lacave & Diez, 2004; Mitchell et al., 1986; Wick & Thompson, 1992).

An expert system consists of a knowledge base encoding the domain knowledge, usually modeled as a set of production rules, a rule interpreter or reasoner that makes use of the knowledge base, and an interface through which the user can query the system for knowledge.

In the literature on expert systems, explanations are mainly understood in one of two ways: an explanation as a line of reasoning, or as a problem-solving activity.

3.1 | Explanations as lines of reasoning

Seeing an explanation as a line of reasoning means mainly understanding it as a trace of the way that production or inference rules are used by the system to make a certain decision. While this kind of explanation mainly accommodates the need of knowledge engineers to understand whether the system's reasoning is technically sound, it (or slight variants of it) can also be provided as an explanation to domain experts (Buchanan & Shortliffe, 1984; Mitchell et al., 1986).

The most famous instantiation of a system that was able to provide this kind of explanation is MYCIN (Buchanan & Shortliffe, 1984). MYCIN is a rule-based system with consultation capabilities developed in the 1970s, created with the aim to provide doctors with diagnostic and therapeutic advice about patients with an infection. MYCIN's expertise consists of a static knowledge base containing domain specific knowledge of an expert, as well as factual knowledge about the particular problem under consideration.

The domain or expert knowledge is modeled by means of production rules (see Table 1), which are used to provide diagnosis solutions to specific cases. That is, the user provides some knowledge about a specific patient as input, and the system uses this knowledge to instantiate rules and to make the diagnosis corresponding to the specific case.

The explanation capability in MYCIN consists of a general question answering module and a reasoning-status checker. The former answers simple English language questions concerning the system's decision in a consultation, or

TABLE 1 Example of a MYCIN rule, in both its logical internal form and English translation

PREMISE:	(AND (SAME CNTXT GRAM GRAMNEG)
	(SAME CNTXT MORPH ROD)
	(SAME CNTXT AIR ANAEROBIC))
ACTION:	(CONCLUDE CNTXT IDENTITY BACTEROIDES TALLY .6)
IF:	(1) The gram stain of the organism is gramneg,
	(2) The morphology of the organism is rod, and
	(3) The aerobicity of the organism is anaerobic
THEN:	There is suggestive evidence (.6) that
	The identity of the organism is bacteroides

about the system's general knowledge. The latter provides explanations about the line of reasoning followed by the system.

The question answering module accepts a set of predefined questions that allow an expert user to query the dynamic knowledge or rationale with respect to the rules, parameters, values, and contexts used in a specific consultation. On the other hand, the reasoning status checker allows the user to go deeper into the sequence of rules that are used. The question answering module accepts two basic explanation commands: a *why* command, by which the user can ascend the reasoning chain and explore higher-goals: and a *how* command, by which the user can descend the chain of inferences exploring how a goal was achieved.

Although the provisioning of explanations as lines of reasoning of why certain decisions were "logically" made improves the interpretability of expert systems, humans, when asked to account for complex reasoning, tend to also reconstruct a story that describes the problem-solving behind the decision. That is, they might reconstruct an explanation that fits their level of knowledge and expertise. For instance, a lay user will not benefit much from a very technical explanation, compared to a domain expert or a knowledge engineer. This is what motivated the reconceptualization of explanations as a problem-solving activity by itself as we will discuss in the next section.

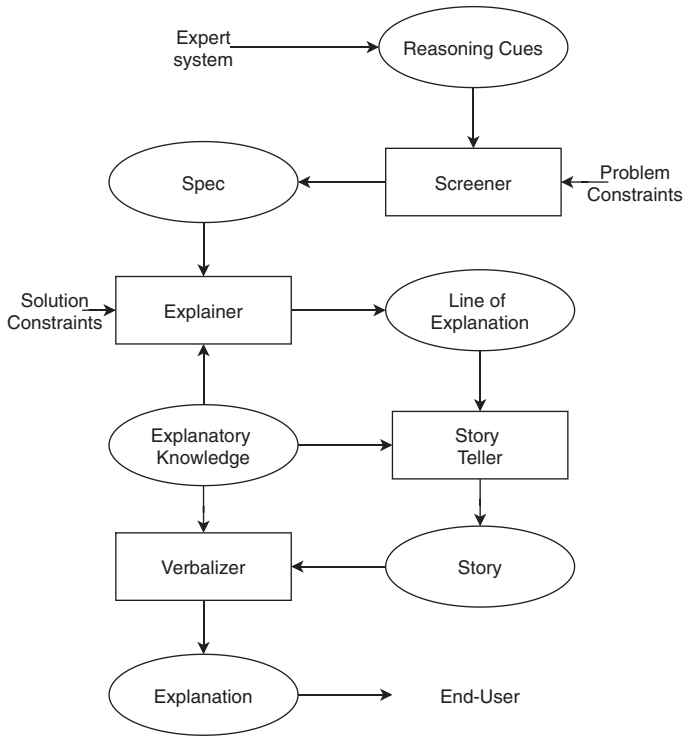
3.2 | Explanations as a problem-solving activity

Conceiving of explanations as problem-solving activities means not only re-constructing the line of reasoning of the system, but also taking into account different levels of abstraction. These could range from very technical to more explanatory explanation formats accommodating different user profiles.

The adaptability of explanations to different types of users can be achieved by de-coupling the explanation capability from the main reasoning functionality, and by focusing the explanation on the problem-solving knowledge used to solve a certain task (Hassling, Clancey, & Runnels, 1984; Wick & Thompson, 1992). An example of an expert system exhibiting this adaptability is Rex (Wick & Thompson, 1992).

Rex was designed to provide explanations of how an expert system moves from the data of a particular case to a final conclusion (a line of explanation) by building a "story" as an abstract of the expert systems reasoning. Rex was an independent component from the expert system used, provided that an interface as well as two knowledge bases existed: a knowledge specification and explanatory knowledge. The former acted as an interface between the knowledge of the expert system and the knowledge of the explanation system, and it covered the problem-solving expertise used to solve problems within the domain. The latter was knowledge used to create an explanation.

The explanation model of Rex is shown in Figure 1. The model takes a set of *reasoning cues*, and a set of *constraints* as input. The reasoning cues consist of knowledge used and inferred by the expert system during the resolution of a certain case. This knowledge is filtered by a set of problem constraints that decide which of these reasoning cues are available to the explanation system. The selected reasoning cues are then mapped to the knowledge specification of the domain, the *screener*. The knowledge specification (*spec*) is the common ground between the expert system and the explanatory system and it is a high-level representation of the domain. It allows the explanation system to abstract from the procedural details of the expert system. The knowledge specification consists of transitions between hypotheses, where any transition requires the satisfaction of some goals and the existence of some reasoning cues. At this step, only some of the transitions might be enabled, thus only some hypotheses can be inferred, and become available to the



CUE uplift
Value : true
Type : direct
Name : the high uplift pressures acting on the dam
Nickname : uplift pressures
Valuename : *cue*

HYPOTHESIS erosion
Value : true
Name : the erosion of soil from under the dam
Nickname : erosion
Valuename : *hypothesis*

GOAL det-cause
Name : determine causal relationships
Nickname : determine causes

CUE SCRIPT erosion-to-sliding
Uses : (<drainage> <uplift> <sliding>)
Supports : <erosion>
Achieves : det-cause
Bottoms : (<drainage> <uplift> <sliding>)
Vconstraint : (and <drainage> <uplift> <sliding>)
Text : <erosion> would cause <broken-pipe> resulting in <drainage> thereby creating <sliding> and eventually <uplift>

GOAL SCRIPT causal
Holds : (<det-cause>)
Text : (simply <det-cause>)

FIGURE 1 Explanation capability as a problem-solving activity (left) and example of explanatory knowledge (right) (Wick & Thompson, 1992)

I attempted to find the cause of an excessive load on a concrete dam. Based on the broken pipes in the foundation, the sliding of the dam, the uplift pressures, and the slow drainage, I was able to make an initial hypothesis. In studying causal relations, I found that the erosion of soil from under the dam would cause broken pipes, resulting in slow drainage, thereby creating increased uplift pressures and eventually sliding of the dam downstream. This led me to conclude erosion was the cause of the excessive load.

FIGURE 2 Example of a line of explanation in Rex (Wick & Thompson, 1992)

explanatory knowledge. The *explainer* can finally build an explanation line by taking into account the knowledge specification and the explanatory knowledge.

The *explanatory knowledge* is a key component of the explanation process (see Figure 1). It models cues, goals, and hypotheses. Transitions among these elements are modeled through scripts. Scripts are represented using a frame-based language. The explainer tries to find an explanation "plan" using only transitions whose hypotheses can be proven. The search of the explanation plan is carried out backward from the final conclusion until reaching the empty hypothesis. Each state in the explanation plan corresponds to an explanation that uses cues and a hypothesis as data, establishes other cues and a hypothesis as conclusions, and traverses certain edges in the knowledge specification.

Once an explanation is found, the *story-teller* organizes it into a consistent flow from data to conclusions. Then, it presents the explanation as a *story* according to a grammar that models the memory structure built during human story-understanding. The basic idea is to extract the information concerning the structure of each hypothesis transition from the line of explanation. Each transition is formatted as a story-tree with a setting, theme, plot, and resolution. The story-tree is then converted to textual description by the *verbalizer* that fills in a template with the problem description, goal description, movement description, and the conclusion of the expert system. A *line of explanation* in Rex looks like the explanation shown in Figure 2.

4 | EXPLANATIONS IN MACHINE LEARNING

While some machine learning models can be considered interpretable by design, namely decision trees, decision rules, and decision tables,[4] the majority of machine learning models work as *black-boxes*. Given an input, a black-box returns

the result of a decision task (classification, prediction, recommendation, etc.), but it does not reveal sufficient details about its internal behavior, resulting in an opaque decision model. For this reason, explainability in machine learning is formulated as the problem of finding an interpretable model that approximates the black-box model as much as possible, typically seeking high *fidelity*.

The literature about explainable or interpretable machine learning is vast. A recent survey on interpretable machine learning methods and techniques can be found in (Guidotti et al., 2018). There, a classification of explanation models is proposed:

- Global methods: The extraction of an explainable counterpart from a black-box model aims at providing an overall approximation of the behavior of the black-box, such that all decisions made by the latter can be tracked in terms of interpretable mechanisms, for example, (Craven & Shaolin, 1995; Frost & Hinton, 2017).
- Local methods: Explanations are built for the decisions made by a black-box model over specific outcomes/instances of a dataset. In this sense, interpretable local models are considered a local approximation of how the black-box works. This kind of explanations can vary greatly depending on the instance considered, for example (Kim, Rodin, & Shah, 2014; Ribeiro, Singh, & Guestrin, 2016b; Ribeiro, Singh, & Guestrin, 2018).
- Introspective methods: Explanations are built by relating inputs to outputs of a black-box model. For instance, explanations can consist of saliency masks for Deep Neural Network models in image classification (such as Convolutional Neural Networks [CNNs]), for example (Hendricks et al., 2016; Park et al., 2016; Same, Wigand, & Müller, 2019) or groups of input–output tokens that are causally related, for example, (Alvarez-Melisa & Jackova, 2017).

In the following, we present some global and local explanation methods: PDPs (partial dependence plots), LIME (local interpretable model-agnostic explanations; Ribeiro et al., 2016b), and SHAP (Shapley Additive explanations; Lundberg & Lee, 2017). Furthermore, we dedicate a section to counterfactual explanations (Mothilal et al., 2020; Watcher, Mittelstadt, & Russell, 2018).

4.1 | Global explanations

The goal of extracting explanations via an interpretable global model is to automatically generate general representations of the black-box model and its relationship to features of the dataset is has been trained on. One possible strategy is to generate symbolic representations of all decisions made by the complex model and represent it in a directly interpretable way. An example of this is the extraction of decision trees, for example (Craven & Shavlik, 1995; Frosst & Hinton, 2017), and decision rules from a trained neural network, for example (Odense & Garcez, 2017; Zhou, Jiang, & Chen, 2003), or the extraction of feature importance vectors, for example, (Lou, Caruana, & Gehrke, 2012; Lou, Caruana, Gehrke, & Hooker, 2013), from noninterpretable models.[5] In some other cases, the interpretable model is a refinement of previous models, which were used to build the black box, such as in the case of Knowledge Neural Networks (Towell & Shavlik, 1993).

A different example can be found in PDPs Friedman (2000), which compute the effect of various variables in the predicted outcome of a machine learning model. This effect can be linear (as in linear regression) or more complex. PDP works by marginalizing the machine learning model output over the distribution of features so that the function shows the relationship between the features one is interested in, and the predicted outcome. PDP works well when one wants to explain two or three features (since it generates 2-D and 3-D plots) and when the features are uncorrelated. In other cases, Accumulated Local Effect plots are used. They work with the conditional instead of the marginal distribution (Apley & Zhu, 2016). Figure 3 shows an example of these explanations.

4.2 | Local explanations

In local explanation methods, the individual predictions of a black-box model can be approximated by generating local surrogate models that are intrinsically interpretable.

This strategy has been implemented for instance in LIME; Ribeiro et al., 2016b). The LIME approach exploits the fact that the trained black-box model can be queried multiple times about the predictions of particular instances. By perturbing the data used for training, LIME generates a new dataset after feeding the black-box model with perturbed

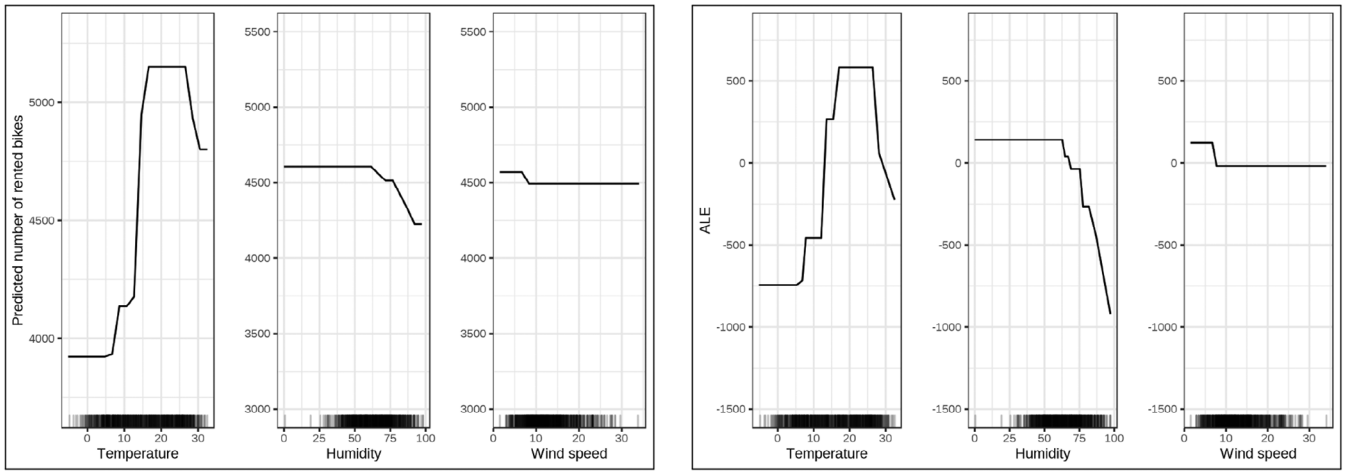


FIGURE 3 Explanations as partial dependence plots—PDPs (left) and Accumulated Local Effect—ALE (right) showing how temperature, humidity, and wind speed affect the predicted number of rented bicycles on a given day (Molnar, 2019). Due to correlation between temperature and humidity, the PDP shows a smaller decrease in predicted number of bikes for high temperature or high humidity compared to the ALE plots. The example shows that when features of a machine learning model are correlated, PDPs are not very accurate and cannot be trusted (Apley & Zhu, 2016)

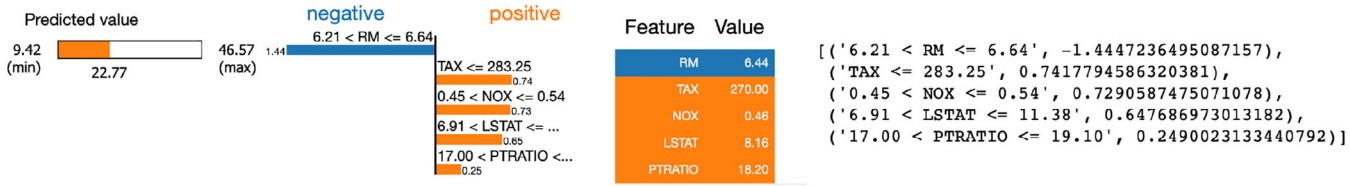


FIGURE 4 Local explanation extracted through LIME in the Boston dataset (Harrison & Rubinfeld, 1978). The dataset contains information collected by the U.S Census Service concerning housing in the area of Boston, Massachusetts. On the left, the median value of owner-occupied homes in $1000's (the predicted value), is explained using a linear regression model using 5 over 14 features (RM, average number of rooms per dwelling; TAX, full-value property-tax rate per $10; 000; NOX, nitric oxides concentration; LSTAT, % lower status of the population; PTRATIO, pupil-teacher ratio by town). On the right, the local explanation in the form of a linear regression using the mentioned features can be appreciated

data and creates a new interpretable model from the predictions made over the new dataset. The local surrogate model is weighted by the proximity of the perturbed instances to the original ones such that it has a high local fidelity.

Methods like LIME generate explanations by creating surrogate models that are interpretable and have a low number of features in order to keep the complexity of the interpretable model low. Figure 4 shows an example of a local explanation extracted by LIME. In the example, the predicted variable is explained using a linear regression.

However, the sampling method used to train the interpretable model is not applicable to situations in which feature spaces are high dimensional or when black-box model decision boundaries are complex. In these scenarios, more features have to be taken into account in order to increase local fidelity, to the detriment of interpretability. An extension of the method, which uses *rules* instead of surrogate models, has recently been proposed by the authors of LIME. The method, called ANCHOR (Ribeiro et al., 2018), uses the same perturbation space as LIME and constructs explanations by adapting their *coverage* to the model structure. In this regard, explanations have a well-defined boundary in terms of their faithfulness to the black-box model.

4.3 | Counterfactual explanations

A counterfactual explanation provides "what-if" information in terms of which alterations of the input features could change the output of a predictive model. A *counterfactual explanation* is then defined as the smallest change to the

"world" (as captured by the input data) that can be made to obtain a desired outcome (Wachter, Mittelstadt, & Russell, 2018); for example, *You were denied a loan because your annual income was £30,000. If your income had been £45,000, you would have been offered a loan*. In this scenario, the borrower receives information regarding why the loan was denied but also will be informed as to what she should do in order to change this outcome.

One of the issues that has to be addressed when generating counterfactuals is that some features might not be changeable (e.g., a person's gender, race, or birth-place). Hence it becomes crucial to present counterfactuals that are indeed *actionable* in the application domain. An actionable example, thus, refers to what can concretely be done next in order to change the outcome of a given decision. For a counterfactual to be actionable it has to meet four properties: proximity, obeying user constraints, sparsity, and causal constraints (Mothilal et al., 2020). Furthermore, presenting users with a set of diversified examples (i.e., a range of suggested actions) can help them shed light on how the system works, and can ease the adoption of these changes.

Unlike explanation methods that depend on approximating the classifier's decision boundary (Ribeiro et al., 2016a), counterfactual explanations have the advantage that they are more human understandable (Mothilal et al., 2020), and that they are always truthful with respect to the underlying model by giving direct outputs of the algorithm (Wachter, Mittelstadt, & Russell, 2018). These properties might prove to be particularly useful in the context of explainability and the GDPR.

An approach to generating counterfactuals was proposed in (Wachter, Mittelstadt, & Russell, 2018). Soon after, the importance of diversity also in counterfactuals was acknowledged by (Russell, 2019), who correspondingly proposed a method to generate diversified counterfactuals for linear models. More recently, Mothilal et al. (2020) proposed Diverse Counterfactual Explanations (DiCE), a novel-model agnostic approach for generating counterfactual examples that are both actionable and diverse.[6] An example of counterfactual explanations using DiCE can be seen in Figure 5.

5 | EXPLANATIONS IN RECOMMENDER SYSTEMS

Recommender systems make use of a large variety of models as back-end engines to serve customized recommendations to users. Such models can be based on Collaborative Filtering, which include Matrix Factorisation (MF; Koren, Bell, & Volinsky, 2009) and all its variants, for example, singular value decomposition (SVD; Nati & Jaakkola, 2003) or nonnegative matrix factorisation (NMF; Lee & Seung, 1999), Nearest Neighbors, and methods based on embeddings such as Deep Learning (Wang, He, Feng, Nie, & Chua, 2018) or Knowledge-based Embeddings for Recommendation (Zhang, Ai, Chen, & Wang, 2018).

Explanations in recommender systems is a popular topic and has received considerable attention in recent years (Nunes & Jannach, 2017; Tintarev & Masthof, 2015). Most of the corresponding work aims to answer the question of why a particular recommendation has been served. This answer can take into account many different aspects used by the recommendation algorithm, such as past interactions characteristics, or contextual information, for example, location of the user, his or her social context, or the time the recommendation is provided.

Query instance (original outcome : 0)

	age	workclass	education	marital_status	occupation	race	gender	hours_per_week	income
0	22.0	Private	HS-grad	Single	Service	White	Female	45.0	0.009411

Diverse Counterfactual set (new outcome : 1)

	age	workclass	education	marital_status	occupation	race	gender	hours_per_week	income
0	57.0	Private	Doctorate	Single	White-Collar	White	Female	45.0	0.724
1	36.0	Private	Prof-school	Married	Service	White	Female	37.0	0.869
2	22.0	Self-Employed	Doctorate	Married	Service	White	Female	45.0	0.755
3	43.0	Private	HS-grad	Married	White-Collar	White	Female	63.0	0.822

FIGURE 5 Example of counterfactual explanations with DiCE (Mothilal et al., 2020). In this example, a neural network was trained to predict the income of a person based on the above eight features (age, work-class, etc.). The first table represents the original query, where the model computed a negative outcome. The second table represents the counterfactual examples

Explainable recommendation systems can be broadly classified into two categories: model-based and post-hoc. The first tackle the mechanistic part of the recommendation, aiming at explaining the way the algorithm proposes a particular recommended item, while the latter analyze the output of a trained recommender in order to infer an explanation for all (recent) recommendations served.

The following subsections provide a brief overview on current state-of-the-art explainable recommender models, and present some forms of explanations meant to increase persuasiveness, effectiveness, efficiency, user satisfaction, and efficiency in platforms that serve recommendations.

5.1 | Explainable recommender system models

One of the most widespread methods which recommendation engines are based on is MF. Factorisation models rely on latent representations of users and items so as to predict either the item(s) with the highest chance to be interacted with, or the rating of an item given by a user. Problems arise when trying to explain the latent factors that contribute to the prediction: the exact meaning of each factor is generally unknown and therefore more information about user interests and item characteristics is required. *Explicit factor models* (EFM; Zhang, 2015) take into account information provided by the user about features of items that she might be interested in (through reviews and explicit feedback) and map them to the latent factors used in the (matrix or tensor) factorization part. Tensor factorization is an extension of EFM's, where the cube user-item-features is used to predict ratings with embedded explanations in terms of features (X. Chen, Qin, Zhang, & Xu, 2016). Implicit feedback for explaining a recommendation has also been proposed by means of neighborhood-based explanations: in these models a recommended item comes with an explanation of the style "*x*% similar users viewed this item," which can be extracted thanks to an explainability regularizer that forces user and item latent vectors to be close if *x*% of users have interacted with the same item.

There are other approaches to explaining recommendations, which are based on the use of external knowledge of items in order to provide personalized explanations on new recommendations. Knowledge-based explanations for recommender systems (Catherine, Mazaitis, Eskénazi, & Cohen, 2017) make use of knowledge graphs that relate item properties and users' behavior in terms of their past interactions with items. With such graphs, different paths can connect a particular user to a particular item (i.e., the graph relationships) in the form of links (either views, purchases, or category), the building blocks of the provided explanations. On a similar note, if user-item relationships are represented as graphs, graph theory can provide insights about how users behave in terms of their interests on different items. For instance, Heckel and Vlachos (2016) proposed a method to compute coclustering to find similar users in terms of their interests and similar items in terms of their properties using an user-item bipartite graph. Explanations can then be retrieved by using shared information between users, considering the purchase/interaction behavior of similar users on recommended items as the core of the explanation.

A different approach to explaining the performance of a recommender engine is to consider it as a black-box that can be probed so as to extract statistical features of recommendations (Peake & Wang, 2018). Explanations can then highlight what percentage of the users have behaved similarly and therefore can provide the confidence on the recommendation to be effective. Besides, the black-box can be approximated by an interpretable version of the recommender engine, for example, association rules or similarity-based models, that can preserve high accuracy while being intrinsically interpretable (Singh & Anand, 2018).

Finally, there recently has been a surge in the number of deep learning-based recommender models deployed in recommender engines (He et al., 2017). Many deep learning techniques, such as CNNs; Seo, Huang, Yang, & Liu, 2017) or Recurrent Neural Networks and Long-Short Term Memory networks (RNN-LSTM) (Hidasi & Karatzoglou, 2018) are used to implement different recommendation strategies, such as sequential recommendations (LSTMs) or context-aware recommendations using user reviews. Attention-based methods are used to highlight the importance of words used in user reviews of past interactions in order to provide explanations about new recommendations. These algorithms use natural language generation in the explanations that can also take into account visual features of the items of interest.

5.2 | Explanation styles in recommender systems

Herlocker, Konstan, and Riedl (2000) compared a large number of different styles of explanations and found that rating histograms generally were users' preferred mechanism for rendering the data behind the recommendations transparent.

Supporting these results, these visual explanations of user styles have proven to be popular in many studies ever since (Bilgic & Mooney, 2005; Cosley, Lam, Albert, Konstan, & Riedl, 2003). Recently, a study using the visual rating histogram paradigm specifically identified user-based explanations and high mean rating values as the most popular styles (Kouki, Schaffer, Pujara, O'Donovan, & Getoor, 2017).

Friedrich and Zanker (2011) proposed a taxonomy to classify different approaches to generate explanations for recommendations. Among the types of explanations in their taxonomy, there are collaborative explanations. These are explanations that justify recommendations based on the amount as well as the concrete values of ratings that derive from similar users, where similarity is typically determined based on similar behavior and preference expressions during past interactions.

The explanation taxonomy proposed by (Papadimitriou, Symeonidis, & Manolopoulos, 2012) extends this classification by making a distinction based on the three fundamental concepts used for explaining recommendations, which are *users*, *items*, and item *features*. They can be used to denote the following explanation styles:

- *User* Style, which provides explanations based on similar users,
- *Item* Style, which is based on choices made by users on similar items, and
- *Feature* Style, which explains the recommendation based on item features (content).

Please note, that any combination of the aforementioned styles is then categorized as a multi-dimensional hybrid explanation style.

For the *User* Style, several collaborative filtering recommender systems, such as the one used by Amazon in their online stores, adopted the following style of justification: "Customers who bought item *X* also bought items *Y*, *Z*, ..." This is called *User* style (Bilgic & Mooney, 2005) as it is based on users performing similar actions like buying or rating items (see also Figure 6). Regarding the *Item* style of explanation, justifications are of the form: "Item *Y* is recommended because you highly rated or bought item *X*, *Z*, ..." Thus, the system depicts those items that is, *X*, *Z*, ..., that mostly influenced the recommendation of item *Y*. Bilgic and Mooney (2005) claimed that the *Item* style is preferable over the *User* style, because it allows users to accurately formulate their true opinion about an item. In case of *Feature* style explanations, the description of items is exploited to determine a match between a current recommended item and observed user interests. For instance, restaurants may be described by features such as location, cuisine, and cost. If a user has demonstrated a preference for Chinese cuisine and Chinese restaurants are recommended, then explanations will note the Chinese cuisine or the restaurants' cost aspects. As part of the work in Coba, Zanker, Rook, & Symeonidis (2018), the authors tested users' preference for different explanation styles in a study. They found that *User* Style explanations were the most preferred. In later studies, they also provided evidence that perception of explanations relates to personality characteristics, and they proposed model-based approaches to further personalize explanations (Coba, Rook, et al., 2019; Coba, Symeonidis, et al., 2019).

6 | EXPLANATIONS IN NEURAL-SYMBOLIC LEARNING AND REASONING

Neural-Symbolic Learning and Reasoning seeks to integrate principles from neural network learning with logical reasoning (Garcez et al., 2015). Although neural networks and symbolic systems are frequently painted as two irreconcilable paradigms, the differences actually are more subtle and less fundamental than frequently presumed.

Symbolic systems operate on the symbolic level where reasoning is performed over abstract, discrete entities following logical rules. A common goal of work on symbolic systems is to model (certain aspects of) common-sense reasoning, for example, the kind of reasoning humans do in their everyday lives, which is considered to automatically allow for

Rating	Number of Neighbours
★	0
★★	0
★★★	0
★★★★	10
★★★★★	23

FIGURE 6 Example of an explanation interface visualizing a User style explanation using the explainability power of nearest neighbors for a target user (Coba, Symeonidis, et al., 2019)

better explainability. Neural networks, on the other hand, operate in the sub-symbolic (or connectionist) level. Individual neurons do not necessarily represent a readily recognizable concept, or any discrete concept at all. Instead, they often model statistical regularities present in the training dataset, imbuing the system with statistical predictive capabilities rather than allowing it to perform sound abstract reasoning. As discussed by Besold, Garcez, Bader, et al. (2017), the integration between both levels could, therefore, bridge low-level information processing such as frequently encountered in perception and pattern recognition with reasoning and explanation on a higher, more cognitive level of abstraction.

Achieving this integration promises a range of benefits such as representations, which are abstract, reusable, and general-purpose. Having these readily available could directly allow to tackle some of the pressing issues with current deep learning practices. While the data efficiency and sample complexity of deep learning systems tend to be very computationally demanding and data-heavy, symbolic approaches are less difficult in that aspect. Furthermore, deep learning approaches often do not generalize well out of the sample distribution and prove to be a limited foundation for transfer learning, whereas symbolic representation can help to overcome these limitations. Last and most importantly in this context, deep learning systems lack transparency while symbolic approaches can be designed in such a way as to follow a humanly comprehensible decision-making process (see, e.g., Garcez et al., 2019; Muggleton, Schmid, Zeller, Tamaddoni-Nezhad, & Besold, 2018).

6.1 | The neural-symbolic integration cycle

Figure 7 illustrates the general idea underlying neural-symbolic approaches. On one side, there is a symbolic system, both writable and readable by human experts. On the other side, we have a neural network capable of taking full advantage of connectionist training methods. The iterative loop between both sides allows for the embedding of symbolic (expert) knowledge into the sub-symbolic model as well as for the extraction of learned and refined knowledge from the connectionist model, which can drive the data-based modification and fine-tuning of predefined rules (see e.g., Besold, Garcez, Stenning, et al., 2017).

This cycle already hints at the four main pillars of neural-symbolic systems: representation, extraction, reasoning, and learning. Knowledge representation provides the mapping between the integrated symbolism and connectionism. The different forms of representations can be divided into rule-based, formula-based, and embeddings. As previously mentioned, the aim is to extract symbolic knowledge given a trained neural network for explaining and reasoning aims. There have also been efforts at integrating neural-symbolic systems into the immediate process of learning. Inductive Logic Programming (ILP), for example, develops a logic program directly from examples (França, Zaverucha, & Garcez, 2014). In addition to this, learning with logical constraints generally has shown to be beneficial for improving the data efficiency (Garnelo & Shanahan, 2019). These constraints can, for example, be integrated as a logic network module on top of a regular neural network. As a consequence, models can further learn relations in-between the inner abstractions as well as guiding the model to explain its prediction. Reasoning is another essential goal of neural-symbolic systems. Successful integration aim to perform symbolic reasoning on the knowledge learned during the training phase (Garcez et al., 2001).

Complementing Deep Learning systems by integrating symbolic representations such as Knowledge Graphs can serve as a lingua franca between humans and AI systems. Sarker, Xie, Doran, Raymer, and Hitzler (2017) propose that methods for explanations should be seen as interactive systems. The authors present a method that enables active

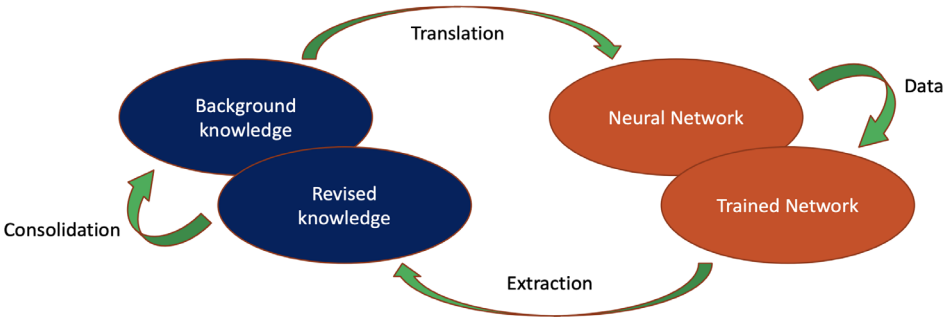


FIGURE 7 Illustration of the neural-symbolic cycle

monitoring for classifiers where humans can act on given explanations. These interactive approaches become increasingly relevant as they provide not only extracted information, but also the ability to act on this information. Against that backdrop symbolic representation are taken to be vital as enablers of human-intelligible explanations. Similarly, Futia and Vetrò (2020) state that hybrid methods will further allow for explanations targeted at nonexperts based on querying and reasoning mechanisms, which are at the core of the integrated semantic components.

6.2 | Explanations via knowledge extraction

Staying close to the neural-symbolic cycle, most traditional approaches to explainability in neural-symbolic systems aim to generate a set of symbolic rules that approximates the behavior of a sub-symbolic model.

The task of generating—usually by via some form of learning—these rules is known as knowledge extraction (Towell & Shavlik, 1993). The extraction process seeks to optimize for different metrics and criteria, namely, accuracy, fidelity, consistency, and comprehensibility. On the one hand, accuracy (i.e., a measure for the performance of the rules on the original test sets) and fidelity metrics (i.e., a measure for the ability of the rules to replicate the behavior of the original sub-symbolic model) relate to performance dimensions of the extracted interpretable model. On the other hand, consistency and comprehensibility are related to the consumer of the rules: rules should be precisely representing the underlying model, but should also be easy to understand and use. This usually requires a trade-off between consistency and comprehensibility.

The extracted rules can then be used to revise and consolidate available background knowledge (often taking the form of domain knowledge). This background knowledge can be used not only to provide meaningful semantics for the explanations—facilitating, in this way, human-machine interactions—but can also be injected back into the sub-symbolic model itself in order to improve its performance (Ziegler et al., 2017).

Returning to the task of knowledge extraction, two main approaches are commonly considered: one of the *decompositional*, the other one *pedagogical*. Algorithms falling of the first type extract rules directly from the structure and weights of the sub-symbolic model. This is usually achieved by first extracting rules that approximate the behavior of each connectionist unit. Then, these unit-level rules are aggregated to form the composite rule base of the neural network as a whole (Andrews et al., 1995). To extract rules in such a way, these methods need access to the internal layers of the sub-symbolic model. Often this access cannot be obtained (e.g., due to intellectual property considerations), but one might still need and want to be able to extract explanations. Algorithms belonging to the class of *pedagogical* approaches overcome this limitation. They treat the sub-symbolic model as an "oracle," and extract information from input–output pairings. A prime example for a pedagogical approach to knowledge extraction is Trepan (Craven & Shavlik, 1995).

Trepan is a tree induction algorithm that recursively extracts decision trees from statistical classifiers, originally intended in particular for use with feed-forward neural networks (but as the original classifier is treated like a generic oracle within the algorithm, Trepan can be considered in principle agnostic to the type of sub-symbolic model at hand). Craven and Shavlik (1995)'s approach can be seen as an extension of the ID2-of-3 algorithm (Murphy & Pazzani, 1991), a method for building decision trees from data based on "m-of-n" rules—that is, m out of n specified conditions must be true to send an example down a particular branch. These tests are usually built by a greedy search algorithm that starts from the single feature that maximizes information gain, and iteratively adds features to the test until information gain is no longer improved by doing so. Trepan combines this with the idea of using a trained machine learning classifier as oracle, in its original version targeting multi-layer perceptrons (MLPs). At each splitting step, the oracle's predicted labels are used instead of the known real labels from the input dataset. Figure 8 shows an example of a Trepan tree extracted from a trained MLP. The use of the classifier as oracle serves two purposes: first, it helps to avoid overfitting to outliers in the training data. Second, and more importantly, it helps to build deeper trees.

While Trepan extracts trees from sub-symbolic models by approximating the models to an arbitrarily close degree without having direct access to their architecture and units, there is still the problem of assessing to what extent the extracted trees are human-understandable. Recent work measured human understandability of decision trees using syntactic and cognitive metrics (Huysmans et al., 2011; Piltaver, Luštrek, Gams, & Martinčić-Ipšić, 2016). Building on these, Confalonieri, Weyde, et al. (2020) also showed how human understandability of surrogate decision trees can be enhanced by using and integrating domain knowledge, for example, in the form of ontologies, in the decision tree extraction.

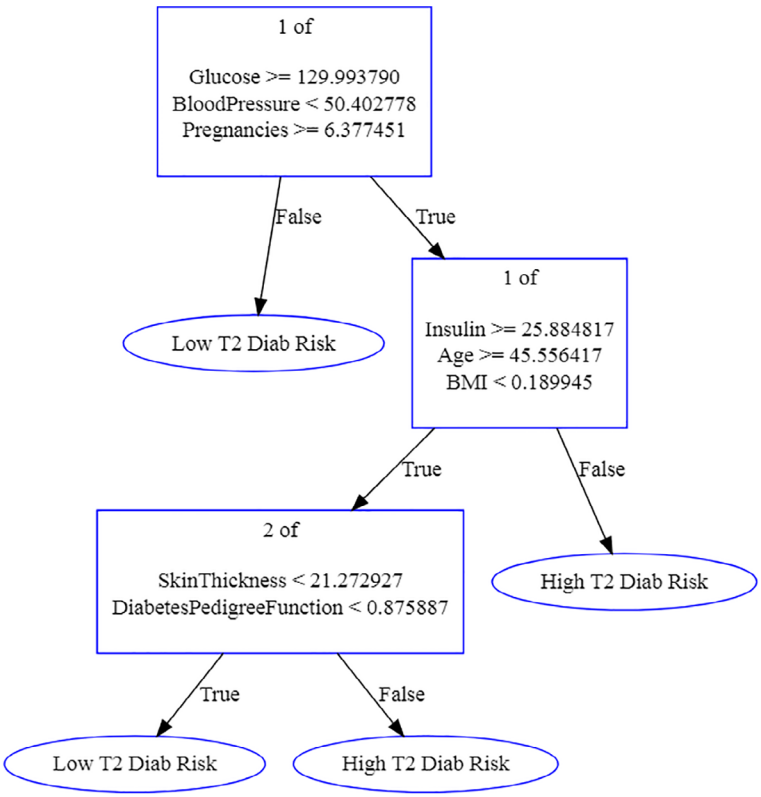


FIGURE 8 Trepan tree extracted from a trained neural network predicting diabetes risk based on the Pima Indians dataset (Smith, Everhart, Dickson, Knowler, & Johannes, 1988)

7 | DISCUSSION

The historical overview provided in this article hints at a categorization of explanations. This categorization relies on the reasoning characteristics of the underlying decision system, namely, symbolic, sub-symbolic, and hybrid.

Expert systems were one of the first realizations of applied AI, where the aim was to build systems able to aid humans in decision-making activities in very specific domains. Making these systems operative required a knowledge acquisition effort in which domain knowledge had to be formally specified. This knowledge formalization was essential to develop intelligent systems able to reason, draw new conclusions, and to generate explanations. Explanations in these systems consisted of either descriptions coupled to the reasoning trace of the system, or descriptions decoupled from the reasoning itself, but more focused on the story behind the decision-making process itself. Since knowledge in expert systems in most cases aimed at modeling (some aspects of) common-sense reasoning, explanations generated by these systems were usually human-understandable. Nonetheless, acquiring and modeling domain knowledge is a complex task, and it is subject to human interpretation and the point of view that the modeler decides to capture.

Machine learning was introduced to alleviate this knowledge acquisition problem. Machine learning algorithms are indeed capable of identifying data patterns from (in most cases) large amounts of data, but this often happens at the price of creating black-box models. An explanation in these systems is mainly understood as an interpretable model that approximates the behavior of the underlying black-box. Explanations of this type allow users to understand why a certain conclusion or recommendation is made, by means of local, global, introspective, or counterfactual explanations. Whereas these explanations seek to maximize metrics such as *accuracy* (i.e., the performance of the extracted interpretable model on the test sets), *fidelity* (i.e., the ability of the extracted interpretable model to replicate the behavior of the black-box model), they also have to be *understandable* by human users. Clearly, accuracy and understandability often compete with each other, and a reasonable trade-off must be found. For instance, a very technical and precise explanation (e.g., in equation form) may be appropriate for a data scientist, but not for a lay person, who prefers possibly a less accurate but more comprehensible representation format of the explanation.

Most explainability methods nowadays are not powerful enough to give guarantees about truthfulness and closeness of the explanation with respect to the underlying model. Most metrics currently in place are lacking a reliable way of expressing this uncertainty. For instance, the measured fidelity is supposed to be a satisfactory proxy of closeness of the representation to the underlying model. However, this metric is limited in its capacity and capability to find

semantically meaningful representations that allow for transparent reasoning, as it is solely optimizing for resemblance of the explained model.

Aspects of understandability of explanations for lay users has for a long time been overlooked. As also pointed out in (Bhatt et al., 2020), the majority of deployments do not focus on the end-users, who are affected by the model, but rather on machine learning engineers, who use explainability to debug the model itself. In practice, there is a gap between explainability and the goal of transparency, since explanations primarily serve "internal" stakeholders rather than "external" ones. To bridge this gap, explanations need to be human-understandable and adaptable to different stakeholders (Ribera & Lapedriza, 2019). Trustworthy systems need to target explanations for different types of user, taking into account their different goals, and providing relevant and selected (customized) information to them. This requires an approach to explainable AI that starts from a user-centered perspective. Related to this, guidelines behind Responsible AI establishing that fairness, accountability and privacy (especially related to data fusion) should be considered when implementing AI models in real environments have been discussed in (Arrieta et al., 2020).

Finally, while explainability has been addressed in some form or another since the mid-1980, its general understanding and definition(s) are still under discussion. In particular, proposing a set of global desiderata for explanations appears to be challenging, since these properties often depend on the application domain. Notwithstanding, we conclude our discussion by pointing out some desiderata that, we believe, should be taken into account for the development of XAI systems, particularly putting the user at the heart of the entire explainability enterprise:

- Causal: Knowing what relationship there is between input and output, or between input features can foster human-understandable explanations. However, causal explanations are largely lacking in the machine learning literature, with only few exceptions such as (Chattopadhyay, Manupriya, Sarkar, & Balasubramanian, 2019). A related problem is then how to measure the causal understanding of an explanation (causability) (Holzinger, Langs, et al., 2019). While this is always possible for explanations of human statements, as the explanation is per-se related to a human model, measuring the causal understanding of an explanation of a machine statement has to be based on a causal model, which is not the case for most machine learning algorithms (Holzinger et al., 2020).
- Counterfactual: Reviewed empirical evidence indicates that humans psychologically prefer counterfactual or contrastive explanations (Miller, 2019). For instance, people do not ask why event *P* happened, but rather why event *P* happened instead of some event *Q*. It is thus important to provide explanations that are both contrastive and direct. Some preliminary steps have been taken in this direction, for example, (Mothilal et al., 2020). Issues related to the diversity and proximity of counterfactuals arise in designing counterfactual explanations.
- Social: Interactive transfer of knowledge is required in which information is tailored according to the recipient's background and level of expertise. Explanations can be conceived of as involving one or more explainers and explainees engaging in information transfer through dialogue, visual representation, or other means (Hilton, 1990). Conversational or argumentative processes can enhance user's inspection of explanations, and increasing user's trust in the system.
- Selective: Explanations do not always need to be complex representations of the real world. They should be epistemically relevant for the explainee. The informational content of explanations has to be selected according to the user's background and needs, as humans do not expect the complete cause of an event. Clearly, this depends on the stakeholders' profiles. For instance, explaining a medical diagnosis to a doctor requires a level of technicality, which, presumably, is not necessary for most lay users.
- Transparent: Explanations should help the explainee in understanding the underlying logic of the decision system, and possibly identifying that the system is wrong. Nonetheless, explanations can sometimes be used to learn about the model or the training data. Therefore, a trade-off between transparency and privacy must be found when generating explanations. Generally, methods to address these concerns will have to be developed for training a differentially private model that is able generate local and global explanations. Harder, Bauer, and Park (2020) is an example of methods of this kind.
- Semantic: If explanations are symbolically grounded—by means of ontologies, conceptual networks, or knowledge graphs—they can support common-sense reasoning. Formal representation and reasoning can in turn enact various forms of knowledge manipulation, such as abstraction and refinement (Confalonieri et al., 2018; Confalonieri, Galliani, et al., 2020; Keet, 2007; Lehmann & Hitzler, 2010; Troquard et al., 2018). These forms of manipulation can play an important role when one wants to develop a system able to provide personalized explanations matching different stakeholder profiles.

- Interactive: Explanations should be interactive, allowing the explainee to revise and consolidate some previous background knowledge. The background knowledge can be used not only to provide meaningful semantics for the explanations, facilitating, in this way, human-machine knowledge interactions, but also injected back to the underlying model to improve its performances (e.g., Kulesza, Burnett, Wong, & Stumpf, 2015).

More generally, if one cares about finding ways of successful communication between humans and AI systems, establishing a common ground of inherent logic from the ground up appears reasonable. This common ground can be facilitated by the modularity that integrates perception at the sub-symbolic level and reasoning at the symbolic level. Recent advancements in AI demonstrate robust solutions for many perception tasks. However, to enforce some understanding of the model at a fundamental level, logical integration using symbolic representations will play an important role in the future.

8 | CONCLUSION

We reviewed the literature on explainability in AI, and provided a historical overview of how the notion of explanation has been conceived from traditional to more recent perspectives, namely in the context of expert systems, of machine learning, of recommender systems, and of neural-symbolic learning and reasoning.

The main goal of this article was not to provide a comprehensive review of the literature on XAI, which can be found in, for example, (Andrews et al., 1995; Arrieta et al., 2020; Fernandez et al., 2019; Guidotti et al., 2018; Mueller et al., 2019; Nunes & Jannach, 2017; Tintarev & Masthof, 2015). We aimed, instead, at describing different notions of explanations, examples thereof, as well as properties, and metrics used to evaluate explanations. The article, thus, contains a wide range of references that the reader can use to "navigate" through different notions of explanations, and gain a deeper understanding of the topic of explainable AI.

In providing this historical overview, we analyzed the different notions of explanation to understand what makes for a good explanation. While we are unable to provide a single answer, one conclusion that can be drawn is that for explanations to be human-understandable, they need to be user-centric explanations. To this end, we proposed some desiderata for explanations, that, in our opinion, are crucial for the development of human-understandable explanations, and, in general, of explainable intelligent systems.

ACKNOWLEDGMENTS

The authors want to thank Daniel Malagarriga for many valuable discussions regarding topics covered in this article. A significant part of the work has been carried out at Alpha Health, Telefónica Innovación Alpha, Barcelona, Spain. The authors thank the Department of Innovation, Research and University of the Autonomous Province of Bozen/Bolzano for covering the Open Access publication costs.

CONFLICT OF INTEREST

The authors have declared no conflict of interest for this article.

AUTHOR CONTRIBUTIONS

Roberto Confalonieri: Conceptualization; investigation; project administration; writing-original draft; writing-review and editing. **Ludovik Coba:** Investigation; writing-original draft; writing-review and editing. **Benedikt Wagner:** Investigation; writing-original draft; writing-review and editing. **Tarek Richard Besold:** Project administration; writing-review and editing.

ORCID

Roberto Confalonieri https://orcid.org/0000-0003-0936-2123
Tarek R. Besold https://orcid.org/0000-0002-8002-0049

ENDNOTES

[1] The right to explanation refers to the right of end-users and, more generally, service consumers, to ask for explanations of why a certain decision was reached by an AI system, such as in the case of loan allowance by a bank, recommendations, and medical diagnosis. For a different point of view on this, please refer to Wachter, Mittelstadt, and Floridi (2017).

[2] Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1.

[3] The categorization of explanations based on the type of system (sub-symbolic, symbolic, and hybrid) also relates to Michie (1988)'s criteria for machine learning: weak, strong and ultra-strong. Michie's aim was to provide operational criteria for various qualities of machine learning that include not only predictive performance but also comprehensibility of learned knowledge. His weak criterion identifies the case in which the machine learner produces improved predictive performance with increasing amounts of data. The strong criterion additionally requires the learning system to provide its hypotheses in symbolic form. Last, the ultra-strong criterion extends the strong criterion by requiring the learner to teach the hypothesis to a human, whose performance is consequently increased to a level beyond that of the human studying the training data alone.

[4] A different problem is then to decide how much these models are human understandable; see for example Huysmans, Dejaeger, Mues, Vanthienen, and Baesens (2011) for a comparison of the comprehensibility of decision tables, trees, and rules.

[5] In Section 6, we will have a closer look at Trepan (Craven & Shavlik, 1995) as a concrete example. Trepan is a global explanation method that extracts decision trees from neural networks. The discussion has been relegated to Section 6 as Trepan can also be considered a neural-symbolic approach.

[6] Here, novelty and diversity are concepts that relate to (serendipitous) information discovery which have been studied, among others, in the fields of information search and recommender systems (Clarke et al., 2008; Vargas & Castells, 2011).

RELATED WIREs ARTICLES

Causability and explainability of artificial intelligence in medicine

REFERENCES

Alvarez-Melis, D., & Jaakkola, T. S. (2017). A causal framework for explaining the predictions of black-box sequence-to-sequence models. *CoRR*, abs/1707.01943.

Andrews, R., Diederich, J., & Tickle, A. B. (1995). Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems*, *8*(6), 373–389.

Apley, D. W., & Zhu, J. (2016). Visualizing the effects of predictor variables in black box supervised learning models. *CoRR*, abs/1612.08468.

Arrieta, A. B., Rodr´ıguez, N. D., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., ... Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115 Retrieved from https://doi.org/10.1016/j.inffus.2019.12.012

Balog, K., & Radlinski, F. (2020). *Measuring Recommendation Explanation Quality: The Conflicting Goals of Explanations*. Proceedings of the 43rd International ACM Sigir Conference on Research and Development in Information Retrieval.

Besold, T. R., Garcez, A. d., Stenning, K., van der Torre, L., & van Lambalgen, M. (2017). Reasoning in non-probabilistic uncertainty: Logic programming and neural-symbolic computing as examples. *Minds and Machines*, *27*(1), 37–77.

Besold, T. R., Garcez, A. S., Bader, S., Bowman, H., Domingos, P. M., Hitzler, P., . . . Zaverucha, G. (2017). Neural-symbolic learning and reasoning: A survey and interpretation. *CoRR*, abs/1711.03902.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., . . . Eckersley, P. (2020). *Explainable Machine Learning in Deployment*. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 648–657. Association for Computing Machinery: New York, NY. Retrieved from https://doi.org/10.1145/3351095.3375624

Bilgic, M., & Mooney, R. J. (2005). *Explaining Recommendations: Satisfaction vs. Promotion*. Proceedings of Beyond Personalization 2005: A Workshop on the Next Stage of Recommender Systems Research at the 2005 International Conference on Intelligent user Interfaces. pp. 13–18.

Buchanan, B. G., & Shortliffe, E. H. (1984). *Rule based expert systems: The MYCIN experiments of the Stanford heuristic programming project*, Boston: Addison-Wesley Longman Publishing Co., Inc.

Catherine, R., Mazaitis, K., Eskénazi, M., & Cohen, W. W. (2017). *Explainable Entity-based Recommendations with Knowledge Graphs*. Proceedings of the Poster Track of the 11th ACM Conference on Recommender Systems (RecSys 2017).

Chattopadhyay, A., Manupriya, P., Sarkar, A., & Balasubramanian, V. N. (2019). *Neural Network Attributions: A Causal Perspective*. In K. Chaudhuri and R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97, pp. 981–990. Long Beach, CA: PMLR. Retrieved from http://proceedings.mlr.press/v97/chattopadhyay19a.html

Chen, L., & Pu, P. (2005). *Trust Building in Recommender Agents*. Proceedings of the Workshop on Web Personalization, Recommender Systems and Intelligent User Interfaces at the 2nd International Conference on e-Business and Telecommunication Networks. pp. 135–145.

Chen, X., Qin, Z., Zhang, Y., & Xu, T. (2016). *Learning to Rank Features for Recommendation Over Multiple Categories*. Proceedings of the 39th International ACM Sigir Conference on Research and Development in Information Retrieval. pp. 305–314. ACM: New York, NY.

Clarke, C. L. A., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., & MacKinnon, I. (2008). *Novelty and Diversity in Information Retrieval Evaluation*. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval—SIGIR'08, 659. doi: https://doi.org/10.1145/1390334.1390446

Coba, L., Rook, L., Zanker, M., & Symeonidis, P. (2019). *Decision Making Strategies Differ in the Presence of Collaborative Explanations*. Proceedings of the 24th International Conference on Intelligent User Interfaces—IUI'19. pp. 291–302. New York, NY: ACM Press. doi: https://doi.org/10.1145/3301275.3302304

Coba, L., Symeonidis, P., & Zanker, M. (2019). Personalised novel and explainable matrix factorisation. *Data Knowledge Engineering*, *122*, 142–158.

Coba, L., Zanker, M., Rook, L., & Symeonidis, P. (2018). *Exploring Users' Perception* of *Collaborative Explanation Styles*. 2018 IEEE 20th Conference on Business Informatics (CBI), pp. 70–78. Retrieved from http://arxiv.org/abs/1805.00977. doi: https://doi.org/10.1109/CBI.2018.00017

Confalonieri, R., Besold, T. R., Weyde, T., Creel, K., Lombrozo, T., Mueller, S. T., & Shafto, P. (2019). *What Makes a Good Explanation? Cognitive Dimensions of Explaining Intelligent Machines*. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), Proceedings of the 41th Annual Meeting of the Cognitive Science Society, CogSci 2019: Creativity + Cognition + Computation. pp. 25–26. Montreal, Canada. cognitivesciencesociety.org. Retrieved from https://mindmodeling.org/cogsci2019/papers/0013/index.html

Confalonieri, R., Eppe, M., Schorlemmer, M., Kutz, O., & Pen˜aloza, R., & Plaza, E. (2018). Upward refinement operators for conceptual blending in the description logic $\mathcal{EL}^{++}$. *Annals of Mathematics and Artificial Intelligence*, *82*(1), 69–99.

Confalonieri, R., Galliani, P., Kutz, O., Porello, D., Righetti, G., & Troquard, N. (2020). *Towards Even More Irresistible Axiom Weakening*. S. Borgwardt & T. Meyer (Eds.). Proceedings of the 33rd International Workshop on Description Logics (DL 2020) Colocated with the 17th International Conference on Principles of Knowledge Representation and Reasoning (KR 2020). Vol. 2663. Rhodes, Greece. CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-2663/paper-8.pdf

Confalonieri, R., Weyde, T., Besold, T. R., & del Prado Martín, F. M. (2020). *Trepan Reloaded: A Knowledge-driven Approach to Explaining Black-box Models*. Proceedings of the 24th European Conference on Artificial Intelligence. Vol. 325, pp. 2457–2464). IOS Press. doi: https://doi.org/10.3233/FAIA200378

Cosley, D., Lam, S. K., Albert, I., Konstan, J. A., & Riedl, J. (2003). *Is Seeing Believing? How Recommender System Interfaces Affect Users' Opinions*. Proceedings of the Conference on Human Factors in Computing Systems (CHI'03). Vol. 5, pp. 585–592.

Craven, M. W., & Shavlik, J. W. (1995). Extracting tree-structured representations of trained networks. In *Neural Information Processing Systems* (pp. 24–30). Cambridge, MA: MIT Press.

Doran, D., Schulz, S., & Besold, T. R. (2017). *What Does Explainable AI Really Mean? A New Conceptualization of Perspectives*. Proceedings of the 1st International Workshop on Comprehensibility and Explanation in AI and ML Colocated with AI*IA 2017 (Vol. *2071*). Available from CEUR-WS.org.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *CoRR*, *abs/1702.08608*.

Doyle, D., Tsymbal, A., & Cunningham, P. (2003). *A review of explanation and explanation in case-based reasoning* (Technical Report). Dublin: Trinity College Dublin, Department of Computer Science.

Fernandez, A., Herrera, F., Cordon, O., Jose del Jesus, M., & Marcelloni, F. (2019). Evolutionary fuzzy Systems for Explainable Artificial Intelligence: Why, when, what for, and where to? *Computational Intelligence Magazine*, *14*(1), 69–81. https://doi.org/10.1109/MCI.2018.2881645

França, M. V., Zaverucha, G., & Garcez, A. S. (2014). Fast relational learning using bottom clause Propositionalization with artificial neural networks. *Machine Learning*, *94*(1), 81–104.

Friedman, J. H. (2000). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*, 1189–1232.

Friedrich, G., & Zanker, M. (2011). A taxonomy for generating explanations in recommender systems. *AI Magazine*, *32*(3), 90. Retrieved from https://aaai.org/ojs/index.php/aimagazine/article/view/2365. https://doi.org/10.1609/aimag.v32i3.2365

Frosst, N., & Hinton, G. E. (2017). *Distilling a Neural Network Into a Soft Decision Tree*. Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 colocated with 16th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2017). CEUR Workshop Proceedings. Vol. 2071.

Futia, G., & Vetrò, A. (2020). On the integration of knowledge graphs into deep learning models for a more comprehensible AI—Three challenges for future research. *Information*, *11*(2), 122. Retrieved from https://doi.org/10.3390/info11020122. https://doi.org/10.3390/info11020122

Garcez, A. S., Besold, T. R., De Raedt, L., Foldiak, P., Hitzler, P., Icard, T., . . . Silver, D. L. (2015). *Neural-symbolic Learning and Reasoning: Contributions and Challenges*. AAAI Spring Symposium—Technical Report.

Garcez, A. S., Broda, K., & Gabbay, D. M. (2001). Symbolic knowledge extraction from trained neural networks: A sound approach. *Artificial Intelligence*, *125*(1–2), 155–207.

Garcez, A. S., Gori, M., Lamb, L. C., Serafini, L., Spranger, M., & Tran, S. N. (2019). Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *CoRR*, abs/1905.06088.

Garnelo, M., & Shanahan, M. (2019). Reconciling deep learning with symbolic artificial intelligence: Representing objects and relations. *Current Opinion in Behavioral Sciences*, *29*, 17–23.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1–42.

Harder, F., Bauer, M., & Park, M. (2020). *Interpretable and Differentially Private Predictions*. The Thirty-fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020. pp. 4083–4090. New York, NY: AAAI Press. Retrieved from https://aaai.org/ojs/index.php/AAAI/article/view/5827

Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, *5*(1), 81–102.

Hasling, D. W., Clancey, W. J., & Rennels, G. (1984). Strategic explanations for a diagnostic consultation system. *International Journal of Man-Machine Studies*, *20*(1), 3–19.

He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T.-S. (2017). *Neural Collaborative Filtering*. Proceedings of the 26th International Conference on World Wide Web. pp. 173–182.

Heckel, R., & Vlachos, M. (2016). Interpretable recommendations via overlapping co-clusters. *CoRR*, *abs/1604.02071*.

Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., & Darrell, T. (2016). Generating visual explanations. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer vision–ECCV 2016* (pp. 3–19). Cham: Springer International Publishing.

Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). *Explaining Collaborative Filtering Recommendations*. Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work—CSCW'00. pp. 241–250.

Hidasi, B., & Karatzoglou, A. (2018). *Recurrent Neural Networks with Top-k Gains for Session-based Recommendations*. Proceedings of the 27th ACM International Conference on Information and Knowledge Management. pp. 843–852. New York, NY: ACM.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, *107*, 65–81.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *CoRR*, abs/1812.04608.

Holzinger, A. (2016). Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Informatics*, *3*, 119–131. Retrieved from http://www.springer.com/computer/ai/journal/40708. https://doi.org/10.1007/s40708-016-0042-6

Holzinger, A., Carrington, A., & Müller, H. (2020). Measuring the quality of explanations: The system causability scale (SCS). *KI—Künstliche Intelligenz (German Journal of Artificial intelligence)*, *34*, 193–198. https://doi.org/10.1007/s13218-020-00636-z

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, *9*(4), e1312 Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1312

Holzinger, A., Plass, M., Kickmeier-Rust, M., Holzinger, K., Crişan, G. C., Pintea, C. M., & Palade, V. (2019). Interactive machine learning: Experimental evidence for the human in the algorithmic loop. *Applied Intelligence*, *49*(7), 2401–2414. https://doi.org/10.1007/s10489-018-1361-5

Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., & Baesens, B. (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, *51*(1), 141–154.

Keet, C. M. (2007). *Enhancing Comprehension of Ontologies and Conceptual Models Through Abstractions*. Proceedings of the 10th Congress of the Italian Association for Artificial Intelligence (ai*ia 2007). pp. 813–821.

Kim, B., Rudin, C., & Shah, J. (2014). *The Bayesian Case Model: A Generative Approach for Case-Based Reasoning and Prototype Classification*. Proceedings of the 27th International Conference on Neural Information Processing Systems. Vol. 2. pp. 1952–1960. Cambridge, MA: MIT Press.

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, *42*(8), 42–49.

Kouki, P., Schaffer, J., Pujara, J., O'Donovan, J., & Getoor, L. (2017). *User Preferences for Hybrid Explanations*. Proceedings of the Eleventh ACM Conference on Recommender Systems—RecSys'17. pp. 84–88.

Kulesza, T., Burnett, M., Wong, W.-K., & Stumpf, S. (2015). *Principles of Explanatory Debugging to Personalize Interactive Machine Learning*. Proceedings of the 20th International Conference on Intelligent User Interfaces. pp. 126–137. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2678025.2701399

Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., & Wong, W.-K. (2013). *Too Much, Too Little, or Just Right? Ways Explanations Impact End Users' Mental Models*. 2013 IEEE Symposium on Visual Languages and Human Centric Computing. pp. 3–10.

Lacave, C., & Diez, F. J. (2004). A review of explanation methods for heuristic expert systems. *The Knowledge Engineering Review*, *19*(2), 133–146. Retrieved from. https://doi.org/10.1017/S0269888904000190

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by nonnegative matrix factorization. *Nature*, *401*, 788–791.

Lehmann, J., & Hitzler, P. (2010). Concept learning in description logics using refinement operators. *Machine Learning*, *78*(1–2), 203–250.

Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, *16*(3), 30:31–30:57.

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, *20*(10), 748–759.

Lou, Y., Caruana, R., & Gehrke, J. (2012). *Intelligible Models for Classification and Regression*. Proceedings of the 18th ACM KDD. pp. 150–158. ACM.

Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). *Accurate Intelligible Models with Pairwise Interactions*. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 623–631. ACM.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, et al. (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 4765–4774). Red Hook, NY: Curran Associates, Inc.

Michie, D. (1988). *Machine Learning in the Next Five Years*. Proceedings of the 3rd European Conference on European Working Session on Learning. pp. 107–122. Marshfield, MA: Pitman Publishing, Inc.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. Retrieved from http://www.sciencedirect.com/science/article/pii/S0004370218305988. https://doi.org/10.1016/j.artint.2018.07.007

Mitchell, T. M., Keller, R. M., & Kedar-Cabelli, S. T. (1986). Explanation-based generalization: A unifying view. *Machine Learning*, *1*(1), 47–80 Retrieved from http://dx.doi.org/10.1023/A:1022691120807

Molnar, C. (2019). Interpretable machine learning. Retrieved from https://christophm.github.io/interpretable-ml-book/

Mothilal, R. K., Sharma, A., & Tan, C. (2020). *Explaining Machine Learning Classifiers Through Diverse Counterfactual Explanations*. FAT* 2020—Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 607–617. New York, NY: Association for Computing Machinery. doi: https://doi.org/10.1145/3351095.3372850

Mueller, S. T., Hoffman, R. R., Clancey, W. J., Emrey, A., & Klein, G. (2019). Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *CoRR*, *abs/1902.01876*. Retrieved from http://arxiv.org/abs/1902.01876

Muggleton, S. H., Schmid, U., Zeller, C., Tamaddoni-Nezhad, A., & Besold, T. (2018). Ultra-strong machine learning: Comprehensibility of programs learned with ilp. *Machine Learning*, *107*(7), 1119–1140.

Murphy, P. M., & Pazzani, M. J. (1991). *ID2-of-3: Constructive Induction of M-of-N Concepts for Discriminators in Decision Trees*. Machine Learning Proceedings 1991.

Nati, N. S., & Jaakkola, T. (2003). *Weighted Low-rank Approximations*. 20th International Conference on Machine Learning. pp. 720–727. AAAI Press.

Nilashi, M., Jannach, D., & bin Ibrahim, O., Esfahani, M. D., & Ahmadi, H. (2016). Recommendation quality, transparency, and website quality for trust-building in recommendation agents. *Electronic Commerce Research and Applications*, *19*, 70–84.

Nunes, I., & Jannach, D. (2017). A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, *27*(3-5), 393–444. Retrieved from http://link.springer.com/10.1007/ s11257-017-9195-0http://ls13-www.cs.tu-dortmund.de/homepage/publications/jannach/Journal UMUAI 2017 2.pdf. https://doi.org/10.1007/s11257-017-9195-0

Odense, S., & Garcez, A. S. (2017). Extracting m of n rules from restricted boltzmann machines. In A. Lintas, S. Rovetta, P. F. Verschure, & A. E. Villa (Eds.), *Artificial neural networks and machine learning—ICANN 2017* (pp. 120–127). Cham: Springer International Publishing.

Papadimitriou, A., Symeonidis, P., & Manolopoulos, Y. (2012). A genralized taxonomy of explanations styles for traditional and social recommender systems. *Data Mining and Knowledge Discovery*, *24*(3), 555–583. Retrieved from https://link.springer.com/content/pdf/10.10072Fs10618-011-0215-0.pdf. https://doi.org/10.1007/s10618-011-0215-0

Park, D. H., Hendricks, L. A., Akata, Z., Schiele, B., Darrell, T., & Rohrbach, M. (2016). Attentive explanations: Justifying decisions and pointing to the evidence. *CoRR*, abs/1612.04757.

Parliament and Council of the European Union. (2016). General data protection regulation.

Peake, G., & Wang, J. (2018. Explanation mining: Post hoc interpretability of latent factor models for recommendation systems. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 2060–2069. New York, NY: Association for Computing Machinery. Retrieved from http://dl.acm.org/doi/10.1145/3219819.3220072. doi: https://doi.org/10.1145/3219819.3220072

Piltaver, R., Luštrek, M., Gams, M., & Martinčić-Ipšić, S. (2016). What makes classification trees comprehensible? *Expert Systems with Applications*, *62* (C, 333–346.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). Model-agnostic interpretability of machine learning. *CoRR*, abs/1606.05386.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). *Why Should I Trust You?: Explaining the Predictions of Any Classifier*. Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144. ACM.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *AAAI* (pp. 1527–1535). New Orleans, Louisiana: AAAI Press.

Ribera, M., & Lapedriza, À. (2019). *Can We Do Better Explanations? A Proposal of User-Centered Explainable AI*. Joint Proceedings of the ACM IUI 2019 Workshops Colocated with the 24th ACM Conference on Intelligent User Interfaces (ACM IUI 2019). Vol. 2327. CEUR-WS.org.

Russell, C. (2019). *Efficient Search for Diverse Coherent Explanations*. Fat* 2019—Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency. pp. 20–28. doi: https://doi.org/10.1145/3287560.3287569

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (2019). *Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models*, Basel, Switzerland: Springer International Publishing.

Sarker, M. K., Xie, N., Doran, D., Raymer, M., & Hitzler, P. (2017). *Explaining Trained Neural Networks with Semantic Web Technologies: First Steps*. Ceur Workshop Proceedings.

Seo, S., Huang, J., Yang, H., & Liu, Y. (2017). *Interpretable Convolutional Neural Networks with Dual Local and Global Attention for Review Rating Prediction*. Proceedings of the Eleventh ACM Conference on Recommender Systems. pp. 297–305. New York, NY: ACM.

Singh, J., & Anand, A. (2018). Posthoc interpretability of learning to rank models using secondary training data. *arXiv*:1806.11330.

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). *Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus*. Proceedings—Annual Symposium on Computer Applications in Medical Care.

Tintarev, N., & Masthof, J. (2015). Explaining recommendations: Design and evaluation. In *Recommender systems handbook* (pp. 217–253). Boston, MA: Springer. https://doi.org/10.1007/978-1-4899-7637-6

Towell, G. G., & Shavlik, J. W. (1993). Extracting refined rules from knowledge-based neural networks. *Machine Learning*, *13*(1), 71–101.

Troquard, N., Confalonieri, R., Galliani, P., Peñaloza, R., Porello, D., & Kutz, O. (2018). *Repairing Ontologies via Axiom Weakening*. In S. A. McIlraith & K. Q. Weinberger (Eds.), Proceedings of the Thirty-second AAAI Conference on Artificial Intelligence, (AAAI-18), pp. 1981–1988. New Orleans, Louisiana. AAAI Press. Retrieved from https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17189

Vargas, S., & Castells, P. (2011). *Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems*. Proceedings of the Fifth ACM Conference on Recommender Systems—RECSYS'11. p. 109. New York, NY: ACM Press. doi: https://doi.org/10.1145/2043932.2043955

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, *7*(2), 76–99. https://doi.org/10.1093/idpl/ipx005

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, *31*(2), 841–887.

Wang, X., He, X., Feng, F., Nie, L., & Chua, T. (2018). *TEM: Tree-Enhanced Embedding Model for Explainable Recommendation*. Proceedings of the 2018 World Wide Web Conference. pp. 1543–1552.

Wick, M. R., & Thompson, W. B. (1992, March). Reconstructive expert system explanation. *Artificial Intelligence*, *54*(1–2), 33–70.

Zhang, Y. (2015). *Incorporating Phrase-level Sentiment Analysis on Textual Reviews for Personalized Recommendation*. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. pp. 435–440. New York, NY: ACM.

Zhang, Y., Ai, Q., Chen, X., & Wang, P. (2018). Learning over knowledge-base embeddings for recommendation. *CoRR*, *abs/1803.06540*.

Zhou, Z.-H., Jiang, Y., & Chen, S.-F. (2003). Extracting symbolic rules from trained neural network ensembles. *AI Communications*, *16*(1), 3–15.

Ziegler, K., Caelen, O., Garchery, M., Granitzer, M., He-Guelton, L., Jurgovsky, J., . . . Zwicklbauer, S. (2017). *Injecting Semantic Background Knowledge into Neural Networks Using Graph Embeddings*. 2017 IEEE 26th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE). pp. 200–205.

How to cite this article: Confalonieri R, Coba L, Wagner B, Besold TR. A historical perspective of explainable Artificial Intelligence. *WIREs Data Mining Knowl Discov*. 2021;11:e1391. https://doi.org/10.1002/widm.1391