# Chapter 2
# Principles of Explainable Artificial Intelligence


**Riccardo Guidotti, Anna Monreale, Dino Pedreschi, and Fosca Giannotti**

## 2.1 Introduction

Artificial Intelligence is nowadays one of the most important scientific and technological areas, with a tremendous socio-economic impact and a pervasive adoption in every field of the modern society. High-profile applications such as medical diagnosis, spam filtering, autonomous vehicles, voice assistants, and image recognition are based on Artificial Intelligence (AI) systems. These AI systems reach their impressive performance mainly through *obscure* machine learning models that "hide" the logic of their internal decision processes to humans because they are not humanly understandable. *Black box models* are tools used by AI to accomplish a task for which either the logic of the decision process is not accessible or it is accessible but not human-understandable. Examples of machine learning black box models adopted by AI systems include neural networks, deep neural networks, ensemble classifiers, SVMs, but also compositions of expert systems, data mining, and hard-coded proprietary software. The choice of using not interpretable machine learning models in AI systems is due to their high performance in terms of accuracy [71]. As a consequence, we have witnessed the rise of a black box society [54], where AI systems adopt obscure decision-making models to carry on their decision processes.

The missing of interpretability on how black box models make decisions and fulfill their tasks is a crucial issue for ethics and a limitation to AI adoption in socially sensitive and safety-critical contexts such as healthcare and law. Also, the problem is not only for lack of transparency but also for possible biases inherited

R. Guidotti (✉) · A. Monreale · D. Pedreschi
University of Pisa, Pisa, Italy
e-mail: riccardo.guidotti@unipi.it; anna.monreale@unipi.it; dino.pedreschi@unipi.it

F. Giannotti
ISTI-CNR Pisa, Pisa, Italy
e-mail: fosca.giannotti@isti.cnr.it



M. Sayed-Mouchaweh (ed.), *Explainable AI Within the Digital Transformation and Cyber Physical Systems*, https://doi.org/10.1007/978-3-030-76409-8_2

by the black boxes from prejudices and artifacts hidden in the training data used by the obscure machine learning models. Indeed, machine learning models are built through a learning phase on training data. These training datasets can contain data coming from the digital traces that people produce while performing daily activities such as purchases, movements, posts in social networks, etc., but also from logs and reports generated in business companies and industries. These "Big Data" can inadvertently contain bias, prejudices, or spurious correlations due to human annotation or the way they are collected and cleaned. Thus, obscure biased models may inherit such biases, possibly causing wrong and unfair decisions. As a consequence, the research in eXplainable AI (XAI) has recently caught much attention [1, 7, 32, 49].

The rest of this chapter is organized as follows. First, Sect. 2.2 shows theoretical, ethical, and legal motivations for the need of an explainable AI. Section 2.3 illustrates the dimensions to distinguish XAI approaches. Then, Sect. 2.4 presents the most common types of explanations and provides some details on the state-of-the-art explanators returning them. Finally, Sect. 2.5 concludes this chapter by discussing practical usability of XAI methods, explanations in real-world applications, and the open research questions.

## 2.2 Motivations for XAI

Why do we need XAI? In the following, we analyze some real cases depicting how and why AI equipped with black box models can be dangerous both for the possibility of discrimination and for the unavailability of justification after incorrect behaviors.

Prejudices and preconceptions on training datasets can be adopted by machine learning classifiers as general rules to be replicated [56]. Automated discrimination is not necessarily due to black box models. In St. George's Hospital Medical School, London, UK, a program for screening job applicants was used during the 1970s and 1980. The program used information from candidates without any reference to ethnicity. However, such a program was found to discriminate against ethnic minorities and women by inferring this information from surnames and place of birth and lowering their chances of being selected for interview [44]. A more recent example is related to Amazon. In 2016, the AI software used by Amazon to determine the areas of the USA to which Amazon would offer free same-day delivery accidentally restricted minority neighborhoods from participating in the program (often when every surrounding neighborhood was allowed).[1] In the same year, the COMPAS score, a predictive model for the "risk of crime recidivism" (proprietary secret of Northpointe), was shown to have a strong ethnic bias from

[1] http://www.techinsider.io/how-algorithms-can-be-racist-2016-4.

the journalists of *propublica.org*.[2] The journalists proved that, according to the COMPAS score, a Black who did not re-offend was classified as "high risk" twice as much as Whites who did not re-offend. On the other hand, White repeat offenders were classified as "low risk" twice as much as Black repeat offenders.

These kinds of biases are tied with the training data. For example, in [15], it is proved that the word embeddings [11] trained on Google News articles exhibit female/male gender stereotypes. Indeed, it is shown that for the analogy "Man is to computer programmer as woman is to $X$," the variable $X$ was replaced by "homemaker" by the trained obscure model. The problem was the literature and texts used to train the model repeating that a woman does the housework. Similarly, in [58], it is shown that a classifier trained to recognize wolves and husky dogs was basing its predictions to distinguish a wolf solely on the presence of snow in the background. This was happening because all the training images with a wolf had snow in the background. These spurious correlations, biases, and implicit rules, hidden in the data, besides discriminating, can also cause wrong and unfair decisions. Unfortunately, in various cases, machine errors could have been avoided if the AI would not have been obscured. In particular, accessing the reasons for the AI decisions is especially crucial in safety-critical AI systems like medicine and self-driving cars, where a possible erroneous outcome could even lead to the death of people. For example, the incident that involved a self-driving Uber car that knocked down and killed a pedestrian in Tempe, Arizona, in 2018.[3] An appropriate XAI method would have helped the company to understand the reasons behind the decision and manage their responsibilities.

Precisely to avoid these cases, the European Parliament turned into law in May 2018 the *General Data Protection Regulation (GDPR)* containing innovative clauses on interpretability for automated decision-making systems. For the first time, the GDPR introduces a *right of explanation* for all individuals to obtain "meaningful explanations of the logic involved" when automated decision-making takes place. Despite conflicting opinions among legal scholars regarding the real scope of these clauses [27, 47, 73], a joint agreement on the need for the implementation of such a principle is crucial, and it is nowadays a big open scientific challenge. Indeed, without a technology able to explain black box models, the right to explanation will remain a "dead letter." How can companies trust their AI products without fully validating and understanding the rationale of their obscure models? And in turn, how can users trust AI services and applications? It would be unthinkable to increase the trust of people and companies in AI without explaining to humans the logic followed by black box models. For these reasons, XAI is at the heart of responsible, open data science across multiple industry sectors and scientific disciplines involving robotics, economics, sociology, and psychology besides computer science.

[2] http://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[3] https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html.

## 2.3 Dimensions of XAI

The goal of XAI is to *interpret* AI reasoning. To *interpret* means to give or provide the meaning or to explain and present in understandable terms some concepts.[4] Therefore, in AI, *interpretability* is defined as the ability to *explain* or to provide the meaning in understandable terms to a human [7, 21]. These definitions assume that the concepts composing an explanation and expressed in the understandable terms are self-contained and do not need further explanations. An explanation is an "interface" between a human and an AI, and it is at the same time both human-understandable and an accurate proxy of the AI.

We can identify a set of dimensions to analyze the interpretability of AI systems that, in turn, reflect on the existing different types of explanations [32]. Some of these dimensions are related to *functional* requirements of explainable Artificial Intelligence, i.e., requirements that identify the algorithmic adequacy of a particular approach for a specific application, while others to the *operational* requirements, i.e., requirements that take into consideration how users interact with an explainable system and what is the expectation. Some dimensions instead derive from the need of *usability* criteria from a user perspective, while others derive from the need of guarantees against any vulnerability issues. Recently, all these requirements have been analyzed [68] to provide a framework allowing the systematic comparison of explainability methods. In particular, in [68], the authors propose *Explainability Fact Sheets* that enable researchers and practitioners to assess capabilities and limitations of a particular explainable method. As an example, given an explanation method $m$, we can consider the following functional requirements. *(i)* Even though $m$ is designed to explain regressors, can we use it to explain probabilistic classifiers? *(ii)* Can we employ $m$ on categorical features even though it only works on numerical features? On the other hand, as an operational requirement, can we consider which is the *function of the explanation*? Provide transparency, assess the fairness, etc.

Besides the detailed requirements illustrated in [68], in the literature, it is recognized as a categorization of explanation methods among fundamental pillars [1, 32]: *(i)* black box explanation vs. explanation by design, *(ii)* global vs. local explanations, and *(iii)* model-specific vs. model-agnostic explainers. In the following, we present details of these distinctions and other important features characterizing XAI methods. Figure 2.1 illustrates a summarized ontology of the taxonomy used to classify XAI methods.

**Black Box Explanation vs. Explanation by Design** We distinguish between black box explanation and explanation by design. In the first case, the idea is to couple an AI with a black box model with an explanation method able to interpret the black box decisions. In the second case, the strategy is to substitute the obscure model with a transparent model in which the decision process is accessible by design. Figure 2.2

[4] https://www.merriam-webster.com/.

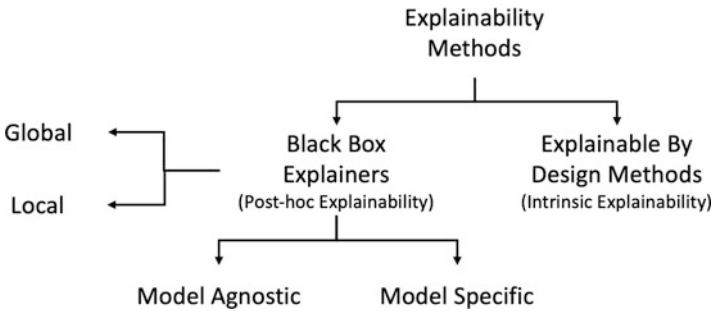


**Fig. 2.1** A summarized ontology of the taxonomy of XAI methods

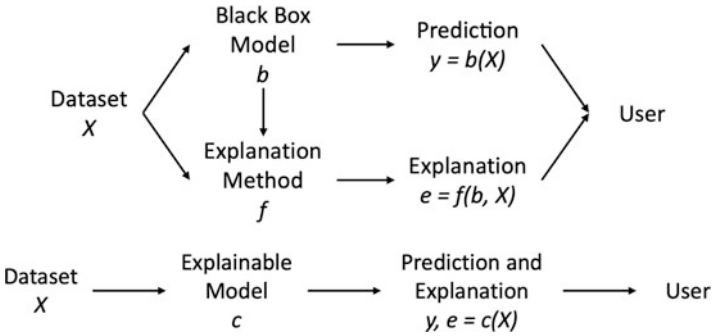


**Fig. 2.2** (Top) Black box explanation pipeline. (Bottom) Explanation by design pipeline

depicts this distinction. Starting from a dataset $X$, the *black box explanation* idea is to maintain the high performance of the obscure model $b$ used by the AI and to use an explanation method $f$ to retrieve an explanation $e$ by reasoning over $b$ and $X$. This kind of approach is the one more addressed nowadays in the XAI research field [20, 45, 58]. On the other hand, the *explanation by design* consists of directly designing a comprehensible model $c$ over the dataset $X$, which is interpretable by design and returns an explanation $e$ besides the prediction $y$. Thus, the idea is to use this transparent model directly into the AI system [61, 62]. In the literature, there are various models recognized to be interpretable. Examples are *decision tree*, *decision rules*, and *linear models* [24]. These models are considered easily understandable and interpretable for humans. However, nearly all of them sacrifice performance in favor of interpretability. In addition, they cannot be applied effectively on data types such as images or text but only on tabular, relational data, i.e., tables.

**Global vs. Local Explanations** We distinguish between global and local explanations depending on whether the explanation allows understanding the whole logic of a model used by an AI system or if the explanation refers to a specific case, i.e., only a single decision is interpretable. A *global* explanation consists in providing a way for interpreting any possible decision of a black box *model*. Generally, the black

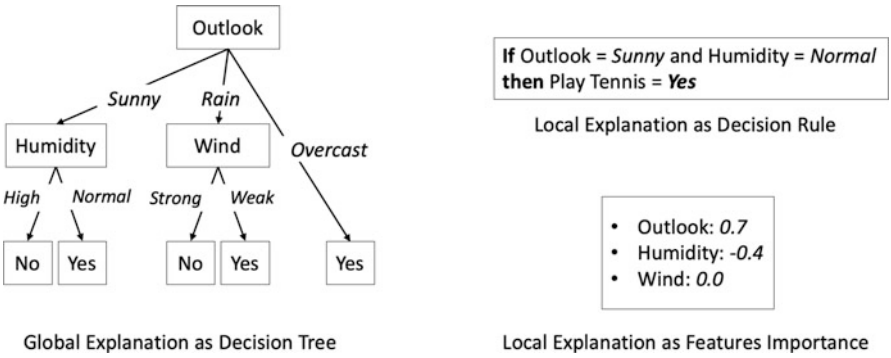


**Fig. 2.3** Explanation examples in the form of decision tree, decision rule, and feature importance

box behavior is approximated with a transparent model trained to mimic the obscure model and also to be human-understandable. In other words, the interpretable model approximating the black box provides a global interpretation. Global explanations are quite difficult to achieve and, up to now, can be provided only for AI working on relational data. A *local* explanation consists in retrieving the reasons for the *outcome* returned by a black box model relatively to the decision for a specific instance. In this case, it is not required to explain the whole logic underlying the AI, but only the reason for the prediction on a specific input instance. Hence, an interpretable model is used to approximate the black box behavior only in the "neighborhood" of the instance analyzed, i.e., with respect only to similar instances. The idea is that in such a neighborhood, it is easier to approximate the AI with a simple and understandable model. Regarding Fig. 2.2 (top), a *global* explanation method $f$ uses many instances $X$ over which the explanation is returned. Figure 2.3 (left) illustrates an example of global explanation $e$ obtained by a decision tree structure for a classifier recommending to play tennis or not. The overall decision logic is captured by the tree that says that the classifier recommends playing tennis or not by first looking at the *Outlook* feature. If its value is *Overcast*, then the prediction is "not to play." Otherwise, if its value is *Sunny*, the classifier checks the *Humidity* feature and recommends "not to play" if the Humidity is *High* and "to play" if it is *Normal*. The same reasoning applies to the other branches of the tree. Still with reference to Fig. 2.2 (top), a *local* explanation method $f$ returns an explanation only for a single instance $x$. Two examples of local explanations are shown in Fig. 2.3 (right). The local rule-based explanation $e$ for a given record $x$ says that the black box $b$ suggested to play tennis because the *Outlook* is *Sunny* and the *Humidity* is *Normal*. On the other hand, the explanation $e$ formed by feature importance says that the black box $b$ suggested playing tennis because the *Outlook* has a large positive contribution, *Humidity* has a consistent negative contribution, and *Wind* has no contribution in the decision.

**Interpretable Models for Explaining AI** To explain obscure AI systems or to replace the black box components, often interpretable models are learned. The

most largely adopted interpretable models are briefly described in the following. A *decision tree* exploits a graph-structure like a tree and composed of internal nodes representing tests on features or attributes (e.g., whether a variable has a value lower than, equal to, or greater than a threshold) and leaf nodes representing a decision. Each branch represents a possible outcome [57]. The paths from the root to the leaves represent the classification rules. The most common rules are *if-then rules*, where the "if" clause is a combination of conditions on the input variables. If the clause is verified, the "then" part reveals the AI action. For a *list of rules*, given an ordered set of rules, the AI returns as the decision the consequent of the first rule that is verified [76]. Finally, *linear models* allow visualizing the *feature importance*: both the sign and the magnitude of the contribution of the attributes for a given prediction [58]. If the sign of an attribute value is positive, then it contributes by increasing the model's output, and otherwise, it decreases it. Higher magnitudes of attribute values indicate a higher influence on the prediction of the model. Examples of such explanations are illustrated in Fig. 2.3.

**User's Desiderata** Since interpretable models are required to retrieve explanations, some desiderata should be taken into account when adopting them [24]. *Interpretability* consists of evaluating to what extent a given explanation is human-understandable. An approach often used for measuring the interpretability is the *complexity* of the interpretable surrogate model. The complexity is generally estimated with a rough approximation related to the *size* of the interpretable model. For example, the complexity of a rule can be measured with the number of clauses in the condition; for linear models, it is possible to count the number of non-zero weights, while for decision trees the depth of the tree.

*Fidelity* consists in evaluating to which extent the interpretable model is able to accurately *imitate*, either globally or locally, the decision of the AI. The fidelity can be practically measured in terms of accuracy score, F1-score, etc. [71] with respect to the decisions taken by the black box model. The fidelity has the goal to determine the soundness and completeness of explanations.

Another important property for the user's point view is the *usability*: an interactive explanation can be more useful than a textual and static explanation. However, machine learning models should also have other ordinary important requirements such as *reliability* [42], *robustness* [34], *causality* [28], *scalability*, and *generality* [55]. Reliability and robustness request that an explanation method should have the ability to maintain certain levels of performance independently from small variations of the parameters or of the input. Causality assumes that controlled changes in the input affect the black box behavior in an expected way, known by the explainer. Generality requires that explanation models are portable to different data (with similar nature) without special constraints or restrictions. Finally, since most of the AI systems need "Big Data," it is opportune to have explainers able to scale to large input data.

Moreover, a fundamental aspect is that every explanation should be *personalized* coherently with the user *background*. Different background knowledge and diverse experiences in various tasks are tied to different notions and requirements for the

usage of explanations. Domain experts can be able to understand complex explanations, while common users require simple and effective clarifications. Indeed, the meaningfulness and usefulness of an explanation depend on the stakeholder [12]. Taking as an example the aforementioned COMPAS case, a specific explanation for a score may make sense to a judge who wants to understand and double-check the suggestion of the AI support system and possibly discover that it biased against Blacks. On the other hand, the same explanation is not useful to a prisoner who cannot change the reality of being Black, while he can find the suggestion meaningful that when he will be older then he would lower his risk down. Moreover, besides these features strictly related to XAI, an interpretable model should satisfy other important general desiderata. For instance, having a high *accuracy* that consists in evaluating to what extent the model accurately takes decisions for unseen instances.

**Model-Specific vs. Model-Agnostic Explainers** We distinguish between model-specific and model-agnostic explanation methods depending on whether the technique adopted to retrieve the explanation acts on a particular model adopted by an AI system or can be used on any type of AI. The most used approach to explain AI black boxes is known as *reverse engineering*. The name comes from the fact that the explanation is retrieved by observing what happens to the output, i.e., the AI decision, when changing the input in a controlled way. An explanation method is *model-specific* or not generalizable [48], if it can be used to interpret only particular types of black box models. For example, if an explanation approach is designed to interpret a random forest [71] and internally use a concept of distance between trees, then such an approach cannot be used to explain the predictions of a neural network. On the other hand, an explanation method is *model-agnostic*, or generalizable, when it can be used independently from the black box model being explained. In other words, the AI's internal characteristics are not exploited to build the interpretable model approximating the black box behavior.

**Time Limitations** The time that the user is allowed to spend on understanding an explanation or is available to do it is a crucial aspect. Obviously, the time availability of a user is strictly related to the scenario where the predictive model has to be used. In some contexts where the user needs to quickly take the decision, e.g., a surgery or an imminent disaster, it is preferable to have an explanation simple and effective. While in contexts where the decision time is not a constraint, such as during a procedure to release a loan, one might prefer a more complex and exhaustive explanation.

**Safety Desiderata** Explainability methods providing interpretable understanding may reveal partial information about the training data, the internal mechanisms of the models, or their parameters and prediction boundaries [14, 65]. Thus, desiderata such as *privacy* [52], *secrecy*, *security*, and *fairness* [56] should be considered to avoid skepticism and increase trust. *Fairness* and *privacy* are fundamental desiderata to guarantee the protection of groups against (direct or indirect) discrimination [60] that the interpretable model does not reveal sensitive information about people [3].

## 2.4 Explanations and Explanators

Increasing research on XAI is bringing to light a wide list of explanations and explanation methods for "opening" black box models. The explanations returned depend on various factors such as the type of task they are needed for and on which type of data the AI system acts, who is the final user of the explanation, if they allow to explain the whole behavior of the AI system (global explanations) or reveal the reasons for the decision only for a specific instance (local explanations). In this section, we review the most used types of explanations and show how some state-of-the-art explanation methods are able to return them. The interest reader can refer to [1, 32] for a complete review of the literature in XAI.

### *2.4.1 Single Tree Approximation*

One of the first approaches introduced to explain neural networks is TREPAN [20]. TREPAN is a global explanation method that is able to model the whole logic of a neural network working on tabular with a single decision tree. The decision tree returned by TREPAN as explanation is a *global transparent surrogate*. Indeed, every path from the root of the tree to a leaf shows the reasons for the final decision reported in the leaf. An example of a decision tree returned by TREPAN is illustrated in Fig. 2.4. This global explanation reveals that the black box first focuses on the value of the feature *rest ECG* and depending on its degree (abnormal, normal, hypertrophy) takes different decisions depending on additional factors such as sex or cholesterol. In particular, TREPAN queries the neural network to induce the decision

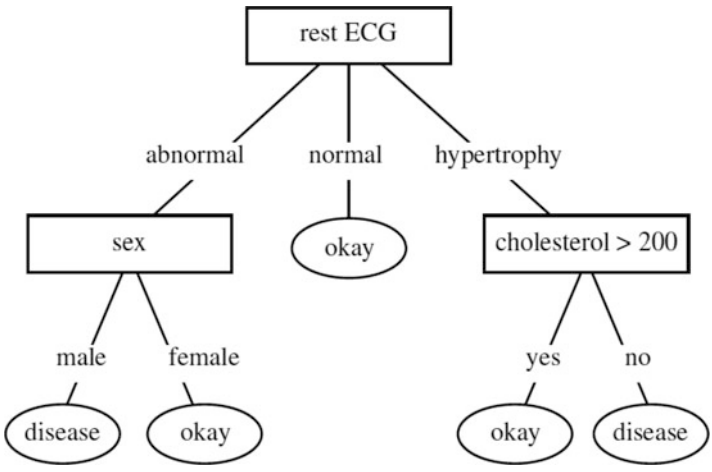


**Fig. 2.4** Example of global tree-based explanation returned by TREPAN

tree by maximizing the gain ratio [71] on the data with respect to the predictions of the neural network. A weakness of common decision trees like ID3 or C4.5 [57] is that the amount of data to find the splits near to the leaves is much lower than those used at the beginning. Thus, in order to retrieve how a neural network works in detail, TREPAN adopts a synthetic generation of data that respect the path of each node before performing the splitting such that the same amount of data is used for every split. In addition, it allows flexibility by using *"m-of-n"* rules where only *m* conditions out of *n* are required to be satisfied to classify a record. Therefore, TREPAN maximizes the fidelity of the single tree explanation with respect to the black box decision. We highlight that even though TREPAN is proposed to explain neural networks, in reality it is model-agnostic because it does not exploit any internal characteristic of neural networks to retrieve the explanation tree. Thus, it can be theoretically employed to explain every type of classifier, i.e., it is model-agnostic.

In [16] is presented an extension of TREPAN that aims to keep the tree explanation simple and compact by introducing four splitting approaches aimed at finding the most relevant features during the tree construction. In [36], genetic programming is used to evolve a single decision tree that approximates the behavior of a neural network ensemble by considering additional genetic features obtained as combinations of the original data and the novel data annotated by the black box models. Both methods described in [16, 36] return explanations in the form of a global decision tree. The readers interested can refer to the papers for more details.

### *2.4.2 Rules List and Rules Set*

A decision rule is generally formed by a set of conditions and by a consequent, e.g., *if conditions, then consequent*. Given a record, a decision rule assigns to the record the outcome specified in the consequent if the conditions are verified [2]. The most common rules are *if-then rules* that take into consideration rules with clauses in conjunction. On the other hand, for *m-of-n* rules given a set of *n* conditions, if *m* of them are verified, then the consequence of the rule is applied [51]. When a set of rules is used, then there are different strategies to select the outcome. For a *list of rules*, the order of the list is considered and the model returns the outcome of the first rule that verifies the conditions [76]. For instance, *falling rule lists* are if-then rules ordered with respect to the probability of a specific outcome [75]. On the other hand, *decision sets* are unordered lists of rules. Basically each rule is an independent classifier that can assign its outcome without regard for any other rule [39]. Voting strategies are used to select the final outcome.

List of rules and set of rules are adopted as explanation both from explanation methods and from transparent classifiers. In both cases, the reference context is tabular data. In [8], the explanation method RXREN unveils with rules list the logic behind a trained neural network. First, RXREN prunes the insignificant input neurons and identifies the data range necessary to classify the given test instance with a

if $((data(I_1) \geq L_{13} \wedge data(I_1) \leq U_{13}) \wedge (data(I_2) \geq L_{23} \wedge data(I_2) \leq U_{23}) \wedge$
$(data(I_3) \geq L_{33} \wedge data(I_3) \leq U_{33}))$ then class $=C_3$
else
if $((data(I_1) \geq L_{11} \wedge data(I_1) \leq U_{11}) \wedge (data(I_3) \geq L_{31} \wedge data(I_3) \leq U_{31}))$
then class $=C_1$
else
class $= C_2$

**Fig. 2.5** Example of the list of rules explanation returned by RXREN

if $(age = 18 - 20)$ and $(sex = male)$ then predict $yes$
else if $(age = 21 - 23)$ and $(priors = 2 - 3)$ then predict $yes$
else if $(priors > 3)$ then predict $yes$
else predict $no$

**Fig. 2.6** Example of the list of rules explanation returned by CORELS

specific class. Second, RXREN generates the classification rules for each class label exploiting the data ranges previously identified and improves the initial list of rules by a process that prunes and updates the rules. Figure 2.5 shows an example of rules returned by RXREN. A survey on techniques extracting rules from neural networks is [4]. All the approaches in [4], including RXREN, are model-specific explainers.

As previously mentioned, an alternative line of research to black box explanation is the design of transparent models for the AI systems. The CORELS method [5] is a technique for building rule lists for discretized tabular datasets. CORELS provides an optimal and certifiable solution in terms of rule lists. An example of rules list returned by CORELS is reported in Fig. 2.6. The rules are read one after the other, and the AI would take the decision of the first rule for which the conditions are verified. Decision sets are built by the method presented in [39]. The if-then rules extracted for each set are accurate, non-overlapping, and independent. Since each rule is independently applicable, decision sets are simple, succinct, and easily to be interpreted. A decision set is extracted by jointly maximizing the interpretability and predictive accuracy by means of a two-step approach using frequent itemset mining and a learning method to select the rules. The method proposed in [63] merges local explanation rules into a unique global weighted rule list by using a scoring system.

### *2.4.3 Partial Dependency*

Another global XAI method for inspecting the behavior of black box models is the partial dependence plot ( PDP). In [32], the black box inspection problem is defined as providing a representation for understanding why the black box returns certain

predictions more likely than others with particular inputs. The PDP visually shows the relationship between the AI decision and the input variables in a reduced feature space clarifying whether the relationship is linear, monotonic, or more complex.

In particular, a PDP shows the marginal effect of a feature on the AI decision [25]. Shortly, a feature is selected and it is varied in its domain. Then, instances are built with values of the selected feature and values from the other features of a given training data. The PDP for a value of the selected feature is the mean probability of classification over the training data or the average regression value. An assumption of the PDP is that the selected feature is not correlated with the other features. Generally, PDP approaches are model-agnostic and used on tabular datasets. In [38], the PROSPECTOR method implementing a PDP is proposed to observe how the output of a black box varies by varying the input changing one variable at a time with an effective way to understand which are the most important features. Figure 2.7 shows the PROSPECTOR PDP for the feature *age* and a black box that predicts the risk of churn. In this example, the PROSPECTOR PDP shows the marginal effect (black line) of the feature *Age* on the predicted outcome *Risk* of a black box classifier. In particular, in this case, the higher is the *Age*, the higher is the probability of *Risk of Churn*. We highlight that for *Age* greater than fifty five this probability markedly increases.

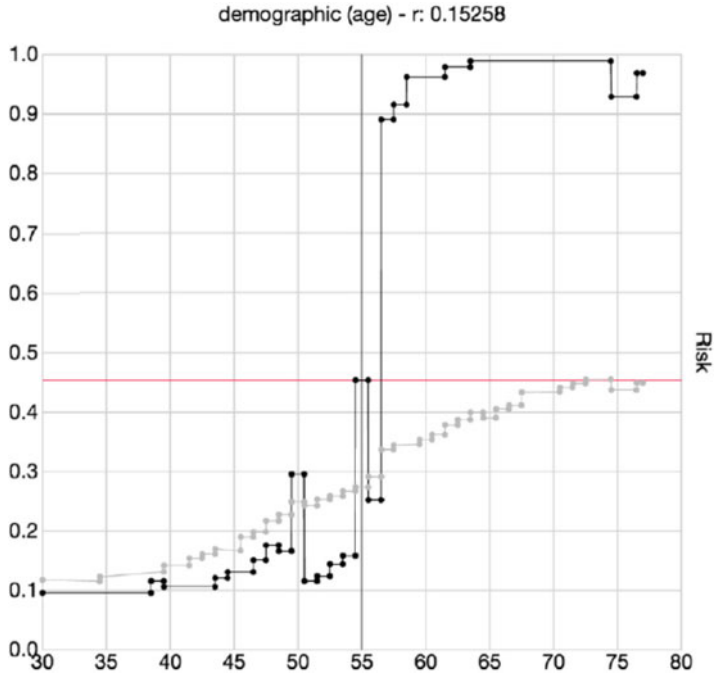


**Fig. 2.7** Example of partial dependence plot

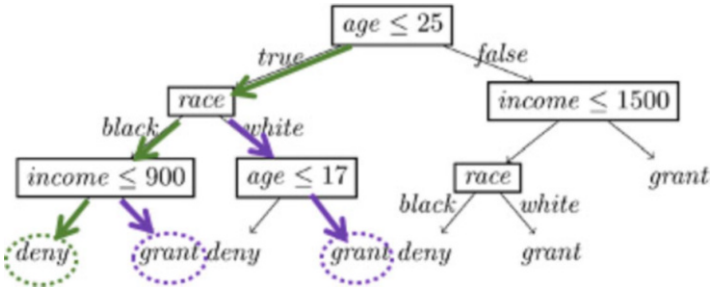


**Fig. 2.8** Example of local factual and counterfactual explanation returned by LORE

### *2.4.4 Local Rule-Based Explanation*

Despite being useful, global explanations can be inaccurate because interpreting a whole model can be complex. On the other hand, even though the overall decision boundary is difficult to explain, locally, in the surrounding of a specific instance, it can be easier. Therefore, a local explanation rule can reveal the *factual reasons* for the decision taken by the black box of an AI system for a specific instance. The LORE method is able to return a local rule-based explanation. LORE builds a local decision tree in the neighborhood of the instance analyzed [30] generated with a genetic procedure to account for both similarity and differences with the instance analyzed. Then, it extracts from the tree a local rule revealing the reasons for the decision on the specific instance (see the green path in Fig. 2.8). For instance, the explanation of LORE for the denied request of a loan from a customer with "age=22, race=black, and income=800" to a bank that uses an AI could be the factual rule *if age* $\leq$ *25 and race* $=$ *black and income* $\leq$*900 then deny*. ANCHOR [59] is another XAI approach for locally explaining black box models with decision rules called anchors. An *anchor* is a set of features with thresholds indicating the values that are fundamental for obtaining a certain decision of the AI. An ANCHOR efficiently explores the black box behavior by generating random instances exploiting a multi-armed bandit formulation.

### *2.4.5 Feature Importance*

A widely adopted form of local explanation, especially for tabular data, consists of *feature importance*. Local explanations can also be returned in the form of feature importance that considers both the sign and the magnitude of the contribution of the features for a given AI decision. If the value of a feature is positive, then it contributes by increasing the model's output; if the sign is negative, then the feature decreases the output of the model. If a feature has a higher contribution than another, then it means that it has a stronger influence on the prediction of the black box

outcome. The feature importance summarizes the decision of the black box model providing the opportunity of quantifying the changes of the black box decision for each test record. Thus, it is possible to identify the features leading to a certain outcome for a specific instance and how much they contributed to the decision.

The LIME model-agnostic local explanation method [58] randomly generates synthetic instances around the record analyzed and then returns the feature importance as the coefficient of a linear regression model adopted as a local surrogate. The synthetic instances are weighted according to their proximity to the instance of interest. The Lasso model is trained to approximate the probability of the decision of the black box in the synthetic neighborhood of the instance analyzed. Figure 2.9 shows the feature importance returned by LIME (central part of the figure) toward the two classes. In this example, the feature *odor=foul* has a positive contribution of 0.26 in the prediction of a mushroom as *poisonous*, *stack-surface-above-ring=silky* has a positive contribution of 0.11, *spore-print-color=chocolate* has a positive contribution of 0.08, *stack-surface-below-ring=silky* has a positive contribution of 0.06, while *gill-size=broad* has a negative contribution of 0.13. Another widely adopted model-agnostic local explanation method is SHAP [45]. SHAP connects game theory with local explanations exploiting the *Shapely values* of a conditional expectation function of the black box to explain the AI. Shapley values are introduced in [64] with a method for assigning "payouts" to "players" depending on their contribution to the "total payout." Players cooperate in a coalition and receive a certain "profit" from this cooperation. The connection with explainability is as follows. The "game" is the decision of the black box for a specific instance. The "profit" is the actual value of the decision for this instance minus the average values for all instances. The "players" are the feature values of the instance that leads toward a certain value, i.e., collaborate to receive the profit. Thus, a Shapley value is the *average marginal contribution* of a feature value across all possible coalitions, i.e., combinations [50]. Therefore, SHAP returns the local unique additive feature importance for each specific record. The higher is a Shapely value, and the higher is the contribution of the feature. Figure 2.10 illustrates an example of SHAP explanation, where the feature importance is expressed in the form of a *force plot*. This explanation shows for each feature the level of the contribution in pushing the black box prediction from the base value (the average model output over the


**Fig. 2.9** Example of explanation based on feature importance by LIME

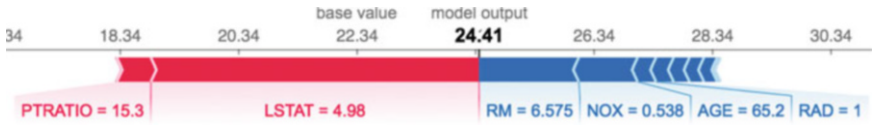


**Fig. 2.10** Example of explanation based on feature importance by SHAP

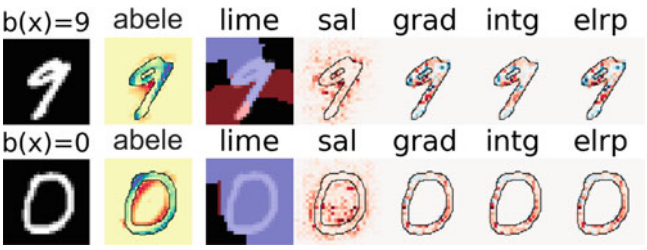


**Fig. 2.11** Example of saliency maps returned by different explanation methods. The first column contains the image analyzed and the label assigned by the black box model *b* of the AI system

dataset, which is 24.41 in this example) to the model output. The features pushing the prediction higher are shown in red; those pushing the prediction lower are shown in blue. Under appropriate settings, LIME and SHAP can also be used to explain the decisions of AI working on textual data and images.

### *2.4.6 Saliency Maps*

The most used type of explanation for explaining AI systems working on images consists of *saliency maps*. A saliency map is an image where each pixel's color represents a value modeling the importance of that pixel for the prediction, i.e., they show the positive (or negative) contribution of each pixel to the black box outcome. Thus, saliency maps are returned by local explanation methods. In the literature, there exist different explanation methods locally explaining deep neural networks for image classification. The two most used model-specific techniques are *gradient attribution methods* like SAL [67], GRAD [66], INTG [69], ELRP [9], and *perturbation-based attribution methods* [23, 77]. Without entering into the details, these XAI approaches aim to assign an importance score to each pixel such that the probability of the deep neural network of labeling the image with a different outcome is minimized, if only the most important pixels are considered. Indeed, the areas retrieved by these methods are also called *attention areas*.

The aforementioned XAI methods are specifically designed for specific DNN models, i.e., they are model-specific. However, under appropriate image transformations that exploit the concept of "superpixels" [58], the model-agnostic explanation methods such as LIME, ANCHOR, and LORE can also be employed to explain

AI working on images for any type of black box model. The attention areas of explanations returned by these methods are tied to the technique used for segmenting the image to explain and to a neighborhood consisting of unrealistic synthetic images with "suppressed" superpixels [29]. On the other hand, the local model-agnostic explanation method ABELE [31] exploits a generative model, i.e., an adversarial autoencoder[46], to produce a realistic synthetic neighborhood that allows retrieving more understandable saliency maps. Indeed, ABELE's saliency maps highlight the contiguous attention areas that can be varied while maintaining the same classification from the black box used by the AI system. Figure 2.11 reports a comparison of saliency maps for the classification of the handwritten digits "9" and "0" for the explanation methods ABELE [31], LIME [58], SAL [67], GRAD [66], INTG [69], and ELRP [9].

### *2.4.7 Prototype-Based Explanations*

Prototype-based explanation methods return as explanation a selection of particular instances of the dataset for locally explaining the behavior of the AI system [50]. Prototypes (or exemplars) make clear to the user the reasons for the AI system's decision. In other words, prototypes are used as a foundation of representation of a category, or a concept [26]. A concept is represented through a specific instance. Prototypes help humans in constructing mental models of the black box model and of the training data used. Prototype-based explainers are generally local methods that can be used independently for tabular data, images, and text. Obviously, prototype-based explanations only make sense if an instance of the data is humanly understandable and makes sense as an explanation. Hence, these methods are particularly useful for images, short texts, and tabular data with few features.

In [13], prototypes are selected as a minimal subset of samples from the training data that can serve as a condensed view of a dataset. Naive approaches for selecting prototypes use the closest neighbors from the training data with respect to a predefined distance function, or the closest centroids returned by a clustering algorithm [71]. In [43], we designed a sophisticated model-specific explanation method that directly encapsulates in a deep neural network architecture an autoencoder and a special prototype layer, where each unit of that layer stores a weight vector that resembles an encoded training input. The autoencoder permits to make comparisons within the latent space and to visualize the learned prototypes such that besides accuracy and reconstruction error, the optimization has a term that ensures that every encoded input is close to at least one prototype. Thus, the distances in the prototype layer are used for the classification such that each prediction comes with an explanation corresponding to the closest prototype. In [18], prototypical parts of images are extracted by a PROTOPNET network, such that each classification is driven by prototypical aspects of a class.

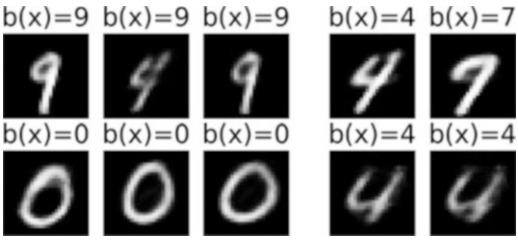


**Fig. 2.12** Example of exemplars (left) and counter-exemplars (right) returned by ABELE. On top of each (counter-)exemplar is reported the label assigned by the black box model *b* of the AI system

Although prototypes are representative of all the data, sometimes they are not enough to provide evidence for the classification without instances that are not well represented by the set of prototypes [50]. These instances are named *criticisms* and help to explain what is *not* captured by prototypes. In order to aid human interpretability, in [37], prototypes and criticisms are selected by means of the *Maximum Mean Discrepancy* (MMD): instances in highly dense areas are good prototypes, and instances that are in regions that are not well explained by the prototypes are selected as criticisms. Finally, the ABELE method [31] enforces the saliency map explanation with a set of exemplar and counter-exemplar images, i.e., images similar to the one under investigation classified for which the same decision is taken by the AI, and images similar to the one explained for which the black box of the AI returns a different decision. The particularity of ABELE is that it does not select exemplars and counter-exemplars from the training set, but it generates them synthetically exploiting an adversarial autoencoder used during the explanation process [40]. An example of exemplars (left) and counter-exemplars (right) is shown in Fig. 2.12.

### *2.4.8 Counterfactual Explanations*

A *counterfactual* explanation shows what should have been different to change the decision of an AI system. Counterfactual explanations are becoming an essential component in XAI methods and applications [6] because they help people in reasoning on the cause–effect relations between analyzed instances and AI decision [17]. Indeed, while direct explanations such as feature importance, decision rules, and saliency maps are important for understanding the reasons for a certain decision, a counterfactual explanation reveals what should be different in a given instance to obtain an alternative decision [73]. Thinking in terms of "counterfactuals" requires the ability to figure a hypothetical causal situation that contradicts the observed one [50]. Thus, the "cause" of the situation under investigation are the features describing the situation and that "caused" a certain decision, while the "event" models the decision.

The most used types of counterfactual explanations are indeed prototype-based counterfactuals. In [74], counterfactual explanations are provided by an explanation method that solves an optimization problem that, given an instance under analysis,

a training dataset, and a black box function, returns an instance similar to the input one and with minimum changes, i.e., minimum distance, but that reverts the black box outcome. The counterfactual explanation describes the smallest change that can be made in that particular case to obtain a different decision from the AI. In [72] is proposed the generation of diverse counterfactuals using mixed integer programming for linear models. As previously mentioned, ABELE [31] also returns synthetic *counter-exemplar images* that highlight the similarities and differences between images leading to the same decision and images leading to other decisions.

Another modeling for counterfactual explanations consists of the logical form that describes a causal situation like: "If $X$ had not occurred, $Y$ would not have occurred" [50]. The local model-agnostic LORE explanation method [30], besides a factual explanation rule, also provides a set of *counterfactual rules* that illustrate the logic used by the AI to obtain a different decision with minimum changes. For example, in Fig. 2.8, the set of counterfactual rules is highlighted in purple and shows that *if income* $>900$ *then grant*, or *if race* $=$ *white then grant*, clarifying which particular changes would revert the decision. In [41] is proposed a local neighborhood generation method based on a Growing Spheres algorithm that can be used for both finding counterfactual instances and acting as a base for extracting counterfactual rules.

## 2.5 Conclusions

This chapter has discussed the problem of interpretability of AI-based decision systems that typically are opaque and hard to understand by humans. In particular, we have analyzed the different dimensions of the problem and the different types of explanations offered by methods proposed by the scientific community. The opportunity to explain complex AI-based systems is fundamental for the diffusion and adoption of those systems in critical domains. One of the most critical ones is the healthcare field where the question of interpretability is far from just intellectual curiosity. The point is that these systems should be used as a support for physicians who have important responsibilities when taking decisions that have a direct impact on health status of humans. For instance, a XAI system, providing details in the form of logical rules or feature importance, could be extremely useful to medical experts who have to monitor and predict the disease evolution of a patient (diabetes detection [70], Alzheimer progression [53], etc.) while understanding the reason for a specific evolution, progression, and complication. Exactly for studying progression and complication, prototype-based explanations and counterfactual explanations can play a crucial role. On the other hand, exemplars and counter-exemplars could be fundamental for identifying brain tumor by comparing with images from magnetic resonance scans [10] and for highlighting through saliency maps the areas of the brain responsible for the decision of the AI system. These are the only examples because there are many other different cases where the knowledge

of the medical staff can be augmented by the knowledge acquired by the machine learning system able to elaborate and analyze myriad of the available information.

Another important field where explainability is applicable is in the context of recommendation systems for getting explainable e-commerce recommendations, explainable social recommendations, and explainable multimedia recommendations. In this context, the goal is to inscribe transparency in the systems but also to provide explanations to final users or system designers who are naturally involved in the loop. In e-commerce, the goal is to explain the ranking of specific recommendations of products [19, 35]. Explainable recommendations also apply to social networks for friend recommendations, recommendation of music, news, travels, tags in images, etc. A useful explanation for recommendation systems could be based on feature importance revealing which are the items contributing positively or negatively to the recommendation. Explainability in social environments is important to increase the users' trustworthiness in the recommendations that is fundamental for the social network sustainability. For instance, in [33], a classifier for predicting the risk of car crash of a driver is equipped with the SHAP explainer that reveals the importance of the features recognizing the risk of collision. Understanding the reasons of recommendations is crucial because it makes the user aware about the technology he/she is using and also about his/her online behavior that enabled the specific recommendation.

Unveiling and interpreting the lending decisions made by an AI-based system is fundamental for the legal point of view and for increasing the social acceptance of these systems. Indeed, these systems based on machine learning models pick up *biases* from the training data. This can lead to learn possible discriminatory behavior against protected groups. In these contexts, interpretability can help in the debugging aimed at detecting those biases and to understand how to have a model able to minimize loan defaults, but also to avoid the discrimination due to certain demographics biases [22]. As a consequence, explainable AI in this setting has a double goal: providing clarification to end user about the reason of the final decisions and providing automated feedback to constantly improve the AI system to eliminate possible ethical issues.

The application domains just discussed are only some of the possible applications of explainable AI. With the AI research advancements, the need of explainability will tend to increase more and more because the complexity of the models could jeopardize their usability. Clearly, the research on explainable AI requires still some effort especially in terms of *personalized and interactive explanations*, i.e., in the study of methods able to provide explanations adaptable to the user background and enabling the human interaction creating the beneficial loop human-machine that could lead the machine to learn from humans and humans from machine.

**Acknowledgments** This work is partially supported by the European Community H2020 programme under the funding schemes: INFRAIA-01-2018-2019 Res. Infr. G.A. 871042 *SoBigData++* (sobigdata.eu), G.A. 952026 *Humane AI Net* (humane-ai.eu), G.A. 825619 *AI4EU* (ai4eu.eu), G.A. 952215 *TAILOR* (tailor.eu), and the ERC-2018-ADG G.A. 834756 "*XAI*: Science and technology for the eXplanation of AI decision making" (xai.eu).

## References


1. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160.
2. Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th Int. Conf. Very Large Data Bases, VLDB* (Vol. 1215, pp. 487–499).
3. Aldeen, Y. A. A. S., Salleh, M., & Razzaque, M. A. (2015). A comprehensive review on privacy preserving data mining. *SpringerPlus, 4*(1), 694.
4. 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.
5. Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2017). Learning certifiably optimal rule lists. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 35–44). ACM.
6. Apicella, A., Isgrò, F., Prevete, R., & Tamburrini, G. (2019). Contrastive explanations to classification systems using sparse dictionaries. In *International Conference on Image Analysis and Processing* (pp. 207–218). Springer.
7. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115.
8. Augasta, M. G., & Kathirvalavakumar, T. (2012). Reverse engineering the neural networks for rule extraction in classification problems. *Neural Processing Letters, 35*(2), 131–150.
9. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One, 10*(7), e0130140.
10. Bakas, S., et al. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge.
11. Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research, 3*, 1137–1155.
12. Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M., & Eckersley, P. (2020). Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 648–657).
13. Bien, J., & Tibshirani, R. (2011). Prototype selection for interpretable classification. *The Annals of Applied Statistics, 5*(4), 2403–2424.
14. Blanco-Justicia, A., Domingo-Ferrer, J., Martínez, S., & Sánchez, D. (2020). Machine learning explainability via microaggregation and shallow decision trees. *Knowledge-Based Systems, 194*, 105532.
15. Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems* (pp. 4349–4357).
16. Boz, O. (2002). Extracting decision trees from trained neural networks. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 456–461).
17. Byrne, R. M. (2019). Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In *IJCAI* (pp. 6276–6282).
18. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems* (pp. 8930–8941).
19. Chen, X., Chen, H., Xu, H., Zhang, Y., Cao, Y., Qin, Z., & Zha, H. (2019). Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In B. Piwowarski, M. Chevalier, É. Gaussier, Y. Maarek, J. Nie & F. Scholer (Eds.), *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21–25, 2019* (pp. 765–774). ACM.

20. Craven, M., & Shavlik, J. W. (1996). Extracting tree-structured representations of trained networks. In *Advances in Neural Information Processing Systems* (pp. 24–30).
21. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
22. Fahner, G. (2018). Developing transparent credit risk scorecards more effectively: An explainable artificial intelligence approach. *Data Anal, 2018*, 17.
23. Fong, R. C., & Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3429–3437).
24. Freitas, A. A. (2014). Comprehensible classification models: A position paper. *ACM SIGKDD Explorations Newsletter, 15*(1), 1–10.
25. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232.
26. Frixione, M., & Lieto, A. (2012). Prototypes vs exemplars in concept representation. In *KEOD* (pp. 226–232).
27. Goodman, B., & Flaxman, S. (2016). EU regulations on algorithmic decision-making and a "right to explanation". In *ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY*. http://arxiv.org/abs/1606.08813v1
28. Goyal, Y., Feder, A., Shalit, U., & Kim, B. (2019). Explaining classifiers with causal concept effect (CACE). arXiv preprint arXiv:1907.07165.
29. Guidotti, R., Monreale, A., & Cariaggi, L. (2019). Investigating neighborhood generation methods for explanations of obscure image classifiers. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 55–68). Springer.
30. Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., & Turini, F. (2019). Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems, 34*(6), 14–23.
31. Guidotti, R., Monreale, A., Matwin, S., & Pedreschi, D. (2019). Black box explanation by learning image exemplars in the latent feature space. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 189–205). Springer.
32. 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 (CSUR), 51*(5), 1–42.
33. Guidotti, R., & Nanni, M. (2020). Crash prediction and risk assessment with individual mobility networks. In *2020 21st IEEE International Conference on Mobile Data Management (MDM)* (pp. 89–98). IEEE.
34. Guidotti, R., & Ruggieri, S. (2019). On the stability of interpretable models. In *2019 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE.
35. He, X., Chen, T., Yen Kan, M., & Chen, X. (2015). TriRank: Review-aware explainable recommendation by modeling aspects.
36. Johansson, U., & Niklasson, L. (2009). Evolving decision trees using oracle guides. In *2009 IEEE Symposium on Computational Intelligence and Data Mining* (pp. 238–244). IEEE.
37. Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems* (pp. 2280–2288).
38. Krause, J., Perer, A., & Ng, K. (2016). Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5686–5697).
39. Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1675–1684). ACM.

40. Lampridis, O., Guidotti, R., & Ruggieri, S. (2020). Explaining sentiment classification with synthetic exemplars and counter-exemplars. In *International Conference on Discovery Science* (pp. 357–373). Springer.
41. Laugel, T., Lesot, M.-J., Marsala, C., Renard, X., & Detyniecki, M. (2017). Inverse classification for comparison-based interpretability in machine learning. arXiv preprint arXiv:1712.08443.
42. Laugel, T., Lesot, M.-J., Marsala, C., Renard, X., & Detyniecki, M. (2019). Unjustified classification regions and counterfactual explanations in machine learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 37–54). Springer.
43. Li, O., Liu, H., Chen, C., & Rudin, C. (2018). Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
44. Lowry, S., & Macpherson, G. (1988). A blot on the profession. *British Medical Journal (Clinical Research Ed.), 296*(6623), 657.
45. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (pp. 4765–4774).
46. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., & Frey, B. (2015). Adversarial autoencoders. arXiv preprint arXiv:1511.05644.
47. Malgieri, G., & Comandé, G. (2017). Why a right to legibility of automated decision-making exists in the General Data Protection Regulation. *International Data Privacy Law, 7*(4), 243–265.
48. Martens, D., Baesens, B., Van Gestel, T., & Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research, 183*(3), 1466–1476.
49. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38.
50. Molnar, C. (2020). *Interpretable machine learning*. Lulu.com
51. Murphy, P. M., & Pazzani, M. J. (1991). Id2-of-3: Constructive induction of m-of-n concepts for discriminators in decision trees. In *Machine learning proceedings 1991* (pp. 183–187). Elsevier.
52. Naretto, F., Pellungrini, R., Monreale, A., Nardini, F. M., & Musolesi, M. (2020). Predicting and explaining privacy risk exposure in mobility data. In *International Conference on Discovery Science* (pp. 403–418). Springer.
53. Oriol, J. D. V., Vallejo, E. E., Estrada, K., Peña, J. G. T., Initiative, A. D. N., et al. (2019). Benchmarking machine learning models for late-onset Alzheimer's disease prediction from genomic data. *BMC Bioinformatics, 20*(1), 1–17.
54. Pasquale, F. (2015). *The black box society*. Harvard University Press.
55. Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., & Turini, F. (2019). Meaningful explanations of black box AI decision systems. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 9780–9784).
56. Pedreshi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 560–568).
57. Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Elsevier.
58. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM.
59. Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
60. Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *Knowledge Engineering Review, 29*(5), 582–638.
61. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.

62. Rudin, C., & Radin, J. (2019). Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review, 1*(2). https://hdsr.mitpress.mit.edu/pub/f9kuryi8
63. Setzu, M., Guidotti, R., Monreale, A., & Turini, F. (2019). Global explanations with local scoring. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 159–171). Springer.
64. Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games, 2*(28), 307–317.
65. Shokri, R., Strobel, M., & Zick, Y. (2019). Privacy risks of explaining machine learning models. CoRR, abs/1907.00164.
66. Shrikumar, A., Greenside, P., Shcherbina, A., & Kundaje, A. (2016). Not just a black box: Learning important features through propagating activation differences. arXiv preprint arXiv:1605.01713.
67. Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.
68. Sokol, K., & Flach, P. A. (2020). Explainability fact sheets: A framework for systematic assessment of explainable approaches. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27–30, 2020* (pp. 56–67). ACM.
69. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. arXiv preprint arXiv:1703.01365.
70. Swapna, G., Vinayakumar, R., & Soman, K. (2018). Diabetes detection using deep learning algorithms. *ICT Express, 4*(4), 243–246.
71. Tan, P.-N. et al. (2006). *Introduction to data mining*. Pearson Education India.
72. Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 10–19).
73. 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.
74. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *HJLT, 31*, 841.
75. Wang, F., & Rudin, C. (2015). Falling rule lists. In *Artificial intelligence and statistics* (pp. 1013–1022).
76. Yin, X., & Han, J. (2003). CPAR: Classification based on predictive association rules. In *Proceedings of the 2003 SIAM International Conference on Data Mining* (pp. 331–335). SIAM.
77. Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European Conference on Computer Vision* (pp. 818–833). Springer.