Data Mining and Knowledge Discovery (2024) 38:3043–3101
https://doi.org/10.1007/s10618-022-00867-8

A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts

Gesina Schwalbe[1,2] · **Bettina Finzel**[2]

Received: 31 March 2021 / Accepted: 9 September 2022 / Published online: 6 January 2023

Abstract

In the meantime, a wide variety of terminologies, motivations, approaches, and evaluation criteria have been developed within the research feld of explainable artifcial intelligence (XAI). With the amount of XAI methods vastly growing, a taxonomy of methods is needed by researchers as well as practitioners: To grasp the breadth of the topic, compare methods, and to select the right XAI method based on traits required by a specifc use-case context. Many taxonomies for XAI methods of varying level of detail and depth can be found in the literature. While they often have a different focus, they also exhibit many points of overlap. This paper unifes these efforts and provides a complete taxonomy of XAI methods with respect to notions present in the current state of research. In a structured literature analysis and meta-study, we identifed and reviewed more than 50 of the most cited and current surveys on XAI methods, metrics, and method traits. After summarizing them in a survey of surveys, we merge terminologies and concepts of the articles into a unifed structured taxonomy. Single concepts therein are illustrated by more than 50 diverse selected example methods in total, which we categorize accordingly. The taxonomy may serve both beginners, researchers, and practitioners as a reference and wide-ranging overview of XAI method traits and aspects. Hence, it provides foundations for targeted, use-case-oriented, and context-sensitive future research.

Responsible editor: Martin Atzmüller, Johannes Fürnkranz, Tomas Kliegr, Ute Schmid.

G. Schwalbe and B. Finzel contributed equally to this work.

✉ Gesina Schwalbe
gesina.schwalbe@continental-corporation.com

Bettina Finzel
bettina.fnzel@uni-bamber g.de

1 Continental AG, Regensburg, Germany

2 Cognitive Systems Group, University of Bamberg, Bamberg, Germany

Keywords Explainable artifcial intelligence · Interpretability · Taxonomy · Meta-analysis · Survey-of-surveys · Review

1 Introduction

Machine learning (ML) models offer the great beneft that they can deal with hardly specifable problems as long as these can be exemplifed by data samples. This has opened up a lot of opportunities for promising automation and assistance systems, like highly automated driving, medical assistance systems, text summaries and question-answer systems, just to name a few. However, many types of models that are automatically learned from data will not only exhibit high performance, but also be black-box—*i.e.*, they hide information on the learning progress, internal representation, and fnal processing in a format not or hardly interpretable by humans.

There are now diverse use-case specifc motivations for allowing humans to *understand* a given software component, *i.e.*, to build up a mental model approximating the algorithm in a certain way. This starts with legal reasons, like the General Data Protection Regulation (Goodman and Flaxman 2017) adopted by the European Union in recent years. Another example are domain specifc standards, like the functional safety standard ISO 26262 (ISO/TC 22/SC 32 2018) requiring accessibility of software components in safety-critical systems. This is even detailed to an explicit requirement for explainability of machine learning based components in the ISO/TR 4804 (ISO/TC 22 Road vehicles 2020) draft standard. Many further reasons of public interest, like fairness or security, as well as business interests like ease of debugging, knowledge retrieval, or appropriate user trust have been identifed (Arrieta et al. 2020; Langer et al. 2021). This requirement to translate behavioral or internal aspects of black-box algorithms into a human interpretable form gives rise to the broad research feld of explainable artifcial intelligence (XAI).

In recent years, the topic of XAI methods has received an exponential boost in research interest (Arrieta et al. 2020; Linardatos et al. 2021; Adadi and Berrada 2018; Zhou et al. 2021). For practical application of XAI in human-AI interaction systems, it is important to ensure a choice of XAI method(s) appropriate for the corresponding use case. Without question, thorough use-case analysis including the main goal and derived requirements is one essential ingredient here (Langer et al. 2021). Nevertheless, we argue that a necessary foundation for choosing correct requirements is a complete knowledge of the different *aspects* (traits, properties) of XAI methods that may infuence their applicability. Well-known aspects are, *e.g.*, portability, *i.e.*, whether the method requires access to the model internals or not, or locality, *i.e.*, whether single predictions of a model are explained or some global properties. As will become clear from our literature analysis in Sect. 5, this only just scratches the surface of aspects of XAI methods that are relevant for practical application.

This paper aims to (1) help beginners in gaining a good initial overview and starting point for a deeper dive, (2) support practitioners seeking a categorization scheme for choosing an appropriate XAI method for their use-case, and (3) to assist researchers in identifying desired combination of aspects that have not or little been considered so far. For this, we provide a complete collection and a structured overview in the form

of a taxonomy of XAI method aspects in Sect. 5, together with method examples for each aspect. The method aspects are obtained from an extensive literature survey on categorization schemes for explainability and interpretability methods, resulting in a meta-study on XAI surveys presented in Sect. 4. Other than similar work, we do not aim to provide a survey on XAI methods. Hence, our taxonomy is not constructed as a means of a suffcient chapter scheme. Instead, we try to compile a taxonomy that is complete with respect to existing valuable work. We believe that this gives a good starting point for an in-depth understanding of sub-topics of XAI research, and research on XAI methods themselves.

Our main contributions are:

- *Complete XAI method taxonomy:* A structured, detailed, and deep taxonomy of XAI method aspects (see Fig. 7); In particular, the taxonomy is complete with respect to application relevant XAI method and evaluation aspects that have so far been considered in literature, according to our structured literature search.
- *Extensive XAI method meta-study:* A survey-of-surveys of more than 50 works on XAI related topics that may serve to pick the right starting point for a deeper dive into XAI and XAI sub-felds. To our knowledge, this is the most extensive and detailed meta-study specifcally on XAI and XAI methods available so far.
- *Broad XAI method review:* A large collection, review, and categorization of more than 50 hand-picked diverse XAI methods (see Table 6). This shows the practical applicability of our taxonomy structure and the identifed method aspects.

The rest of the paper reads as follows: In Sect. 2 we reference some related work, recapitulate major milestones in the history of XAI and provide important defnitions of terms. This is meant for those readers less familiar with the topic of XAI. After that, we detail our systematic review approach in Sect. 3. The results of the review are split into the following chapters: The detailed review of the selected surveys is presented in Sect. 4, including their value for different audiences and research focus; and Sect. 5 details collected XAI method aspects and our proposal of a taxonomy thereof. Each considered aspect is accompanied by illustrative example methods, a summary of which can be found in Table 6. We conclude our work in Sect. 6.

2 Background

This chapter gives some background regarding related work (Sect. 2.1), and the history of XAI including its main milestones (Sect. 2.2). Lastly, Sect. 2.3 introduces some basic terms and defnitions used throughout this work. Experienced readers may skip the respective subsections.

2.1 Related work

Meta-studies on XAI methods

Short meta-studies collecting surveys of XAI methods are often contained in the related work section of XAI reviews like (Linardatos et al. 2021, Sec. 2.3). These are, however, by nature restricted in length and usually concentrate on most relevant

and cited reference works like Gilpin et al. (2018), Adadi and Berrada (2018), Arrieta et al. (2020). We here instead consider a very broad and extensive collection of surveys. By now, also few dedicated meta-studies can be found in literature. One is the comprehensive and recent systematic literature survey conducted by Vilone and Longo (2020, Chap. 4). This reviews 53 survey articles related to XAI topics, which are classifed differentially according to their focus. Their literature analysis reaches further back in time and targets even more general topics in XAI than ours. Hence, several of their included surveys are very specifc, quite short, and do not provide any structured notions of taxonomies. Also, the focus of their review lies in research topics, while the slightly more detailed survey-of-surveys presented here also takes into account aspects that relate to suitability for beginners and practitioners. Another dedicated survey-of-surveys is the one by Chatzimparmpas et al. (2020). With 18 surveys on visual XAI this is smaller compared to ours, and has quite a different focus: They review the intersection of XAI with visual analytics. Lastly, due to their publication date they only include surveys from 2018 or earlier, which misses on important recent work as will be seen in Fig. 2. The same holds for the very detailed DARPA report by Mueller et al. (2019) from 2019 Hence, both papers miss many recent works covered in our meta-study (cf. Fig. 2). Further, the DARPA report has a very broad focus, also including sibling research felds related to XAI. And, similar to Vilone et al., they target researchers, hence give no estimation of what work is especially suitable for beginners or practitioners. Most recently, Saeed and Omlin (2021) conducted a meta-survey of 58 XAI related papers, but with the focus of identifying current challenges in the feld.

(Meta-)studies on XAI method traits

Related work on taxonomies is mostly shallow, focused on a sub-feld, or is purely functional in the sense that taxonomies merely serve as a survey chapter structure. A quite detailed, but less structured collection of XAI method traits can be found in the courses of discussion in Murdoch et al. (2019), Carvalho et al. (2019), Burkart and Huber (2021), Mueller et al. (2019), Li et al. (2020). The focus of each will be discussed later in Sect. 4. But despite their detail, we found that each article features unique aspects that are not included in the others. The only systematic meta-review on XAI method traits known to us is the mentioned survey by Vilone and Longo (2020). They, however, generally review notions related to XAI, not concretely XAI method traits. Only a shallow taxonomy is derived from parts of the retrieved notions (cf. Vilone and Longo 2020, Sec. 5).

Use-case analysis

A related, wide and important feld which is out-of-scope of this paper is that of use-case and requirements analysis. This, *e.g.*, includes analysis of the background of the explanation receiver (Miller 2019). Instead, we here concentrate on fnding aspects of the XAI methods themselves that may be used for requirements analysis and formulation, *e.g.*, whether the explainability method must be model-agnostic or the amount of information conveyed to the receiver must be small. For further detail on use-case specifc aspects the reader may be referred to one of the following surveys further discussed in Sect. 4 (Miller 2019; Guidotti et al. 2018; Gleicher 2016; Langer et al. 2021; Arrieta et al. 2020).

2.2 History of XAI

XAI is not a new topic, although the number of papers published in recent years might suggest it. The abbreviation XAI for the term explainable artifcial intelligence was frst used by van Lent et al. (2004) (see Carvalho et al. 2019, Sec. 2.3). According to Belle (2017), the frst mention of the underlying concept already dates back to the year 1958, when McCarthy described his idea of how to realize AI systems. In his seminal paper he promoted a declarative approach, based on a formal language and a problem-solving algorithm that operates on data represented in the given formal language. Such an approach could be understood by humans and the system's behavior could be adapted, if necessary (McCarthy 1958). Basically, McCarthy described the idea of inherently transparent systems that would be explainable by design.

Inherently transparent approaches paved the way for expert systems that were designed to support human decision-makers (see Jackson 1998). Due to the big challenge to integrate, often implicit, expert knowledge, these systems lost their popularity in the 90's of the last century. Meanwhile, deep neural networks (DNNs) have become a new and powerful approach to solve sub-symbolic problems. The frst approaches aiming at making neural network decisions transparent for debugging purposes date back to the mid 90's. For example in 1992, Craven and Shavlik presented several methods to visualize numerical data, such as decision surfaces (Craven and Shavlik 1992).

Due to the introduction of the new European General Data Protection Regulation (GDPR) in May 2018, transparency of complex and opaque approaches, such as neural networks, took on a new meaning. Researchers and companies started to develop new AI frameworks, putting more emphasis on the aspect of accountability and the "right of explanation" (Goodman and Flaxman 2017; Council 2017). Besides debugging and making decisions transparent to developers or end-users, decisions of a system now also had to be comprehensible for further stakeholders. According to Adadi and Berrada (2018) the term XAI gained popularity in the research community after the Defense Advanced Research Projects Agency (DARPA) published its paper about explainable artifcial intelligence, see Gunning and Aha (2019).

In their defnition, XAI efforts aim for two main goals. The frst one is to create machine learning techniques that produce models that can be explained (their decision-making process as well as the output), while maintaining a high level of learning performance. The second goal is to convey a user-centric approach, in order to enable humans to understand their artifcial counterparts. As a consequence, XAI aims for increasing the trust in learned models and to allow for an effcient partnership between human and artifcial agents (Gunning and Aha 2019). In order to reach the frst goal, DARPA proposes three strategies: deep explanation, interpretable models and model induction, which are defned in Sect. 2.3. Among the most prominent XAI methods that implement this goal for deep learning, especially in computer vision, are for example LIME (Ribeiro et al. 2016), LRP (Bach et al. 2015) and Grad-CAM (Selvaraju et al. 2017). The second, more user-centric, goal defned by DARPA requires a highly inter-disciplinary perspective. This is based on felds such as computer science, social sciences as well as psychology in order to produce more explainable models,

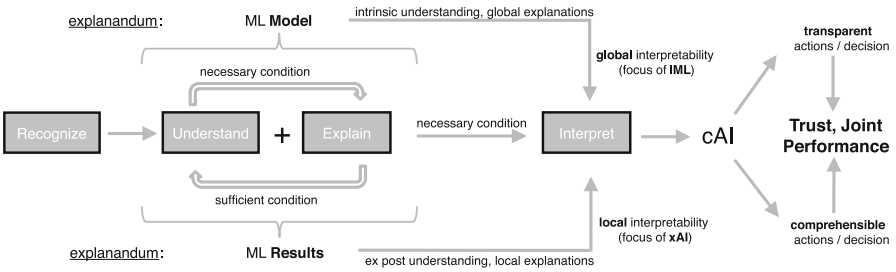

Fig. 1 A framework for comprehensible artifcial intelligence (Bruckert et al. 2020)

suitable explanation interfaces, and to communicate explanations effectively under consideration of psychological aspects.

Miller (2019), *e.g.*, made an important contribution to the implementation of the user-centric perspective with his paper on artifcial intelligence from the viewpoint of the social sciences. He considers philosophical, psychological, and interaction-relevant aspects of explanation, examining different frameworks and requirements. An important turn towards a user-centric focus of XAI was also supported by Rudin (2019) in her paper from 2019, where she argues for using inherently interpretable models instead of opaque models, motivated by the right to explanation and the societal impact of intransparent models.

Another milestone in the development of XAI is the turn toward evaluation metrics for explanations (Mueller et al. 2021). The XAI community now acknowledges more in depth that it is not enough to generate explanations, but that it is also crucial to evaluate how good they are with respect to some formalized measure.

2.3 Basic definitions

This section introduces some basic terms related to XAI which are used throughout this paper. Detailed defnitions of the identifed XAI method aspects will be given in Sect. 5.

Important work that is concerned with defnitions for XAI can be found in, *e.g.*, Lipton (2018), Adadi and Berrada (2018), and the work of Doshi-Velez and Kim (2017) who are often cited as base work for formalizing XAI terms [cf. *e.g.* (Linardatos et al. 2021)]. Some defnitions are taken from Bruckert et al. (2020), who present explanation as a process involving recognition, understanding and explicability respectively explainability, as well as interpretability. The goal of the process is to achieve making an AI system's actions and decisions transparent as well as comprehensible. An overview is given in Fig. 1.

In the following we will assume we are given an AI system that should (partly) be explained to a human. The system encompasses one (or several) AI models and any pre- and post-processing. We use the following nomenclature:

Understanding is described as the human ability to recognize correlations, as well as the context of a problem and is a necessary precondition for explanations (Bruckert et al. 2020). The concept of understanding can be divided into mechanistic under-

standing (*"How does something work?"*) and functional understanding (*"What is its purpose?"*) (Páez 2019).

Explicability refers to making properties of an AI model inspectable (Bruckert et al. 2020).

Explainability goes one step further than *explicability* and aims for making (a) the context of an AI system's reasoning, (b) the model, or (c) the evidence for a decision output accessible, such that they can be *understood* by a human (Bruckert et al. 2020).

Transparency is fulflled by an AI model, if its algorithmic behavior with respect to decision outputs or processes can be *understood* by a human *mechanistically* (Páez 2019). Transparency will be discussed more closely in Sect. 5.1.2.

Explaining means utilizing *explicability* or *explainability* to allow a human to *understand* a model and its purpose (Bruckert et al. 2020; Páez 2019).

Global explanations *explain* the model and its logic as a whole ("How was the conclusion derived?") (Adadi and Berrada 2018).

Local explanations *explain* individual decisions or predictions of a model ("Why was this example classifed as a car?") (Adadi and Berrada 2018).

Interpretability means that an AI model's decision can be *explained globally* or *locally* (with respect to *mechanistic understanding*), and that the model's purpose can be *understood* by a human actor (Páez 2019)(*i.e. functional understanding*).

Correctability means that an AI system can be adapted by a human actor in a targeted manner in order to ensure correct decisions (Kulesza et al. 2015; Teso and Kersting 2019; Schmid and Finzel 2020). Adaptation refers either to re-labelling of data (Teso and Kersting 2019) or to changing of a model by constraining the learning process (Schmid and Finzel 2020).

Interactivity applies if one of the following is possible: (a) interactive explanations, meaning a human actor can incrementally explore the internal working of a model and the reasons behind its decision outcome; or (b) the human actor may adapt the AI system (*correctability*).

Comprehensibility relies, similar to *interpretability*, on local and global *explanations* and *functional understanding*. Additionally, *comprehensible* artifcial intelligence fulflls *interactivity* (Bruckert et al. 2020; Schmid and Finzel 2020). Both, *interpretable* presentation and intervention are considered as important aspects for in depth *understanding* and therefore preconditions to *comprehensibility* (see also Gleicher 2016).

Human-AI system is a system that contains both algorithmic components and a human actor, which have to cooperate to achieve a goal (Schmid and Finzel 2020). We here consider in particular **explanation systems**, *i.e.*, such human-AI systems in which the cooperation involves *explanations* about an algorithmic part of the system (the *explanandum*) by an *explanator* component, to the human interaction partner (the *explainee*) resulting in an action of the human (Bruckert et al. 2020).

Explanandum (*what is to be explained*, cf. Sect. 5.1) refers to what is to be *explained* in an *explanation system*. This usually encompasses a model (*e.g.*, a deep neural network). We here also refer to an explanandum as the object of explanation.

Explanator (*the one that explains*, cf. Sect. 5.2; also called explainer) is the *explanation system* component providing *explanations*.

Explainee *(the one to whom the explanandum is explained)* is the receiver of the *explanations* in the *explanation system*. Note that this often but not necessarily is a human. *Explanations* may also be used, *e.g.*, in multi-agent systems for communication between the agents and without a human in the loop in most of the information exchange scenarios.

Interpretable models are defned as machine learning techniques that learn more structured representations, or that allow for tracing causal relationships. They are *inherently interpretable* (cf. defnition in Sect. 5.2), *i.e.*, no additional methods need to be applied to *explain them*, unless the structured representations or relationship are too complex to be processed by a human actor at hand.

Interpretable machine learning (iML) is the area of research concerned with the creation of *interpretable* AI systems (*interpretable models*).

Model induction (also called model distillation, student-teacher approach, or reprojection (Gleicher 2016)) is a strategy that summarizes techniques which are used to infer an approximate *explainable* model—the (*explainable*) *proxy* or *surrogate model*—by observing the input-output behavior of a model that is *explained*.

Deep explanation refers to combining deep learning with other methods in order to create hybrid systems that produce richer representations of what a deep neural network has learned, and that enable extraction of underlying semantic concepts (Gunning and Aha 2019).

Comprehensible artificial intelligence (cAI) is the result of a process that unites *local interpretability* based on *XAI* methods and *global interpretability* with the help of *iML* (Bruckert et al. 2020). The ultimate goal of such systems would be to reach *ultra-strong machine learning*, where machine learning helps humans to improve in their tasks. For example, (Muggleton et al. 2018) examined the *comprehensibility* of programs learned with Inductive Logic Programming, and (Schmid et al. 2016; Schmid and Finzel 2020) showed that the *comprehensibility* of such programs could help lay persons and experts to *understand* how and why a certain prediction was derived.

Explainable artificial intelligence (XAI) is the area of research concerned with *explaining* an AI system's decision.

3 Approach to literature search

The goals of this paper are to (1) provide a complete overview of relevant aspects or properties of XAI methods, and (2) ease fnding the right survey providing further details. In order to achieve this, a systematic and broad literature analysis was conducted on papers in the time range of 2010–2021. The target of this meta-survey are works that either contain reviews on XAI methods, or considerations on XAI metrics and taxonomy aspects.

Table 1 Main used search phrases for the search for XAI taxonomies in the Google Scholar database with approximate number of matches

Matches	Search phrase
> 300	Explain AI taxonomy
ca. 80	XAI taxonomy toolbox guide
> 300	Explainable AI taxonomy toolbox guide
ca. 20	Explain interpret AI artifcial intelligence DNN Deep Learning ML machine learning taxonomy framework toolbox guide XAI

3.1 Search

Our search consisted of two iterations, one directly searching for work on XAI taxonomies, and one more general search for general XAI surveys.

Work on XAI taxonomies
The iteration for work on XAI taxonomies started with an initial pilot phase. Here we identifed common terms associated directly with XAI taxonomies (for abbreviations both the abbreviation and the full expression must be considered):

- machine learning terms: AI, DNN, Deep Learning, ML
- explainability terms: XAI, explain, interpret
- terms associated with taxonomies: taxonomy, framework, toolbox, guide

In the second search phase, we collected Google Scholar[1] search results for combinations of these terms. The main considered search terms are summarized in Table 1. For each search, the frst 300 search results were scanned by the title for relevance to our search target. Promising ones then were scanned by abstract. Only work that we could access was fnally included. In the last phase, we scanned the works recursively for references to further relevant reviews.

General XAI surveys
The second iteration collected search results for XAI surveys that do not necessarily propose a taxonomy, but possibly implicitly use one. For this, we again conducted the mentioned three search phases. The search terms now were the more general ones "XAI" and "XAI survey". These again were scanned frst by title, then by abstract. This resulted in a similar number of fnally chosen and in-depth assessed papers as the taxonomy search (not counting duplicate search results).

Lastly, we also included surveys and toolboxes that were additionally pointed out by the reviewers.

3.2 Categorization

For the sub-selection and categorization, we considered the general focus, the length, level of detail, target audience, citation count per year, and recency.

[1] https://scholar.google.com.

General focus Regarding the *general focus*, we sorted the obtained reviews into three categories:

General XAI method collections (Sect. 4.2): Reviews that contain a broad collection of XAI methods without an explicit focus on a specifc sub-feld;
Domain-specific XAI method collections (Sect. 4.3): Reviews that also contain a collection of XAI methods, but with a concrete focus on application domain (*e.g.* medicine), method technologies (*e.g.* inductive logic programming), or method traits (*e.g.* black-box methods);
Conceptual reviews (Sect. 4.1): Reviews that do not contain or not focus on XAI method examples, but on conceptual aspects for the feld of XAI like taxonomy or metrics;
Toolboxes (Sect. 4.4): Summary of a complete toolbox with implementation.

Length For the *length* we considered the number of pages up to the bibliography, resulting in four categories (cf. Fig. 2): short (up to 6 pages), medium (up to 15 pages), long (up to 50 pages), and very long (more than 50 pages).

Target audience As *target audiences* we considered three types: beginners in the feld of XAI (potentially coming from a different domain), practitioners, and researchers. A review was associated with one or several target audiences, if it specifcally targeted that audience, or if it was judged practically suitable or interesting for readers of that audience.

Citation count per year The *citation count per year* of the surveys was used as a proxy for reception. It was collected from the popular citation databases Google Scholar, Semantic Scholar,[2] OpenCitations,[3] and NASA ADS.[4] The highest result (mostly google scholar) was chosen for comparison.

Recency was compared via the publication year.

Using these categorizations, some further works were ruled out for inclusion into the survey of surveys. Excluded were reviews which: would not suffciently match our search target; were too specifc (*e.g.* comparison of only very few specifc methods); are very short (up to 4 pages) and are covered by prior or successive work of the authors, or are not often cited or not suffciently focused.

3.3 Results

We reviewed over 70 surveys on XAI with publication date up to beginning of 2021 Most of them are from the years 2019 to 2021 (cf. Fig. 2), which well fts the exponential increase in XAI methods that was observed so far (Arrieta et al. 2020; Linardatos et al. 2021; Adadi and Berrada 2018; Zhou et al. 2021). These were analyzed for XAI method aspects, taxonomy structuring proposals, and suitable example methods for each aspect. In the following, a sub-selection of more than 50 surveys are detailed that were most cited or of general interest, as well as some examples of toolboxes.

[2] https://www.semanticscholar.org/.

[3] https://opencitations.net/.

[4] https://ui.adsabs.harvard.edu/.

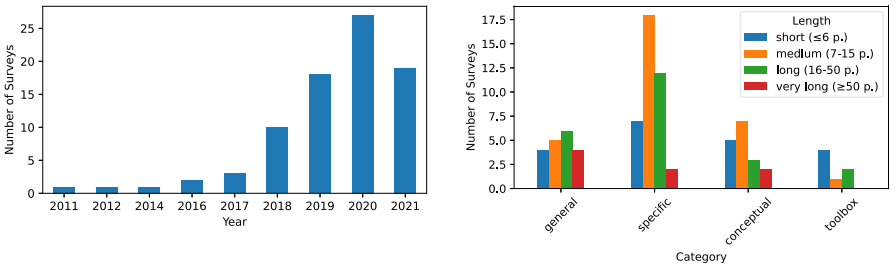

Fig. 2 Distribution of the reviewed surveys over the years 2010 to 2021 (*left*), and distribution of the survey lengths by general focus category (*right*)

To exemplify the aspects of our proposed taxonomy, we selected again more than 50 concrete XAI methods that are reviewed in example sections for the corresponding XAI aspects. The selection focused on high diversity and recency of the methods, in order to establish a broad view on the XAI topic. Finally, each of the methods were analyzed for main taxonomy aspects, which is summarized in Table 6. During the fnalization phase of this paper we came across further interesting surveys (Guidotti 2022; Zhang et al. 2021), and the base work (Tintarev and Masthoff 2007) on explanations in recommender systems. We are happy to honorably mention these here, but we do not analyze the works in detail in the following.

4 A survey of surveys on XAI methods and aspects

Along with the vastly growing demand and interest in XAI methods came an abundance of helpful reviews on the topic. This encompasses overviews of all kinds of aspects, methods, and metrics, each with custom focus. Given this battering amount of literature, it may be hard, especially for beginners and practitioners, to fnd a survey suited to their needs. The needs can encompass a specifc focus, a desired level of detail and length, or others. This chapter provides a short survey of XAI surveys that shall help a reader in fnding appropriate material to dig further into the topic of XAI.

The focus of the literature search lay on works that use, propose, or deduce a structured view on XAI methods and metrics. Search and categorization criteria are detailed in Sect. 3. In the rest of this section, the more than 50 selected surveys are reviewed with respect to their key focus points, level of detail, and their target audience. While this survey of surveys for sure will not be complete, it should both give a good overview on the breadth of XAI, as well as serve as a good starting point when looking for a deeper dive into topics of the feld.

Our results so far promise that many more helpful surveys on the topic of XAI will come up within the next years (cf. Fig. 2). A key result of this structured meta-study will be presented later in Sect. 5: The works are analyzed for taxonomy aspects of XAI methods, resulting in the—to our knowledge—most complete taxonomy of XAI methods available so far.

Table 2 Overview on clusters of reviewed conceptual surveys on XAI

Focus	
Broad	Gunning et al. (2019), Lipton (2018), Mueller et al. (2019), Guidotti et al. (2021)
Stakeholder perspective	Gleicher (2016), Langer et al. (2021)
HCI perspective	Miller (2019), Chromik and Schüßler (2020), Ferreira and Monteiro (2020), Mueller et al. (2021)
XAI method evaluation	Doshi-Velez and Kim (2017), Zhou et al. (2021)

4.1 Conceptual reviews

By now, many works have gathered and formulated important concepts related to XAI research. We here roughly divided the available literature by their main focus: broadly focused surveys, surveys from the perspectives of stakeholders and human-computer interaction, and fnally surveys with an explicit focus on XAI metrics. An overview can be found in Table 2.

Broad conceptual surveys
A very short, high-level, and beginner-friendly introduction to XAI can be found in the work of Gunning et al. (2019). They derive four open challenges for the feld: user-centric explanations, tradeoff between accuracy and interpretability, automated abstraction, and appropriate end-user trust. Regarding an XAI method taxonomy, a base for many later surveys was by Lipton (2018). In this medium-length work, Lipton provides a small and high-level XAI taxonomy. The focus, namely motivation and desiderata for interpretability, are held broad and suitable for beginners and interdisciplinary discussion. In contrast, the very broad and long DARPA report by Mueller et al. (2019) provided later in 2019 is targeted at researchers of the feld. The goal of the report is to broadly capture all relevant developments and topics in the feld of XAI. This resulted in a detailed meta-study on state-of-the-literature, and a detailed list of XAI method aspects and metrics (chapters 7 and 8). More recently, Guidotti et al. (2021) illustrate some key dimensions to distinguish XAI approaches in a beginner-friendly book section. They present a broad collection of most common explanation types and state-of-the-art explanators respectively, and discuss their usability and applicability. This results in a conceptual review of taxonomy aspects and evaluation criteria.

XAI from a stakeholder perspective
The early work by Gleicher (2016) highlights the stakeholder perspective on XAI problems. In the medium-length survey, Gleicher suggests a framework of general considerations for practically tackling an explainability problem. A similar focus is set by Langer et al. (2021) in their long recent work from 2021. They review in detail XAI from the point of view of satisfying stakeholder desiderata. For this they collect standard goals of XAI problems and propagate that XAI design choices should take into account all stakeholders in an interdisciplinary manner.

XAI from a HCI perspective

When humans interact with AI-driven machines, this human-machine-system can beneft from explanations obtained by XAI. Hence, there are by now several surveys concentrating on XAI against the background of human-computer-interaction (HCI). An important and well-received base work in this direction is that of Miller (2019). He conducts a long, detailed, and extensive survey of work from philosophy, psychology, and cognitive science regarding explanations and explainability. The main conclusions for XAI research are: (a) (local) explanations should be understood contrastively, *i.e.*, they should clarify why an action was taken instead of another; (b) explanations are selected in a biased manner, *i.e.*, do not represent the complete causal chain but few selected causes; (c) causal links are more helpful to humans than probabilities and statistics; and (d) explanations are social in the sense that the background of the explanation receiver matters. A much shorter paper was provided by Chromik and Schüßler (2020). They rigorously develop a taxonomy for evaluating black-box XAI methods with the help of human subjects. Furthermore, they make some concrete suggestions for study design. The related survey by Ferreira and Monteiro (2020) is slightly longer, and may serve as an entry point to the topic for researchers. They present a taxonomy of XAI that links with computer science and HCI communities, with a structured and dense collection of many related reviews. Very recent work on XAI metrics is provided by Müeller et al. in their 2021 paper (Mueller et al. 2021). They collect concrete and practical design principles for XAI in human-machine-systems. Several relevant XAI metrics are recapitulated, and their broad collection of related surveys may serve as an entry point to the research feld of human-AI-systems.

Surveys with a focus on evaluation

Continuing on the HCI perspective, a hot topic in the feld of XAI are metrics for measuring the quality of explanations for human receivers. An early and base work on XAI metric categorization is by Doshi-Velez and Kim (2017). The medium-length survey collects latent dimensions of interpretability with recommendations on how to choose and evaluate an XAI method. Their taxonomy for XAI metrics is adapted by us, classifying metrics into human grounded, functionally grounded, and application grounded ones (cf. Sect. 5.3). A full focus on metrics for evaluating XAI methods is set in the recent work by Zhou et al. (2021). This medium-length, detailed meta-study reviews diverse XAI metrics, aligned with shallow taxonomies both for methods and metrics.

4.2 Broad method collections

There by now exists an abundance of surveys containing broad collections of XAI methods, serving as helpful starting points for fnding the right method. The following shortly highlights surveys that feature a generally broad focus (for specifcally focused surveys see Sect. 4.3). This summary shall help to fnd a good starting point for diving deeper into methods. Hence, the surveys are frst sorted by their length (as a proxy for the amount of information), and then by their reception (citations per year). The latter was found to strongly correlate with age. An overview of the discussed surveys is given in Table 3.

Table 3 Overview on focus points/specialties and audience (Aud.) of discussed general XAI method collections

	Aud.	Focus/specialty	Detail
Short			
Došilović et al. (2018)	B, R	XAI for supervised learning	1
Biran and Cotton (2017)	R	Local explanations and interpretable models	1
Benchekroun et al. (2020)	B, P	Industry point of view	5
Medium			
Gilpin et al. (2018)	R	XAI research aspects	1
Du et al. (2019)	B	Perturbation based attention visualization	3
Murdoch et al. (2019)	B, P	Predictive, descriptive, relevant desiderata for explanations, practical examples	5
Goebel et al. (2018)		Need for XAI	3
Islam et al. (2021)	B, R	Demonstration of XAI methods in case-study (credit scoring)	5
Long			
Adadi and Berrada (2018)	P	XAI research landscape (methods, metrics, cognitive aspects, human-machine-systems)	1
Carvalho et al. (2019)	B, R	Aspects associated with XAI (esp. motivation, properties, desirables)	2
Xie et al. (2020)	B	Explanation of DNNs	3
Das and Rad (2020)	R	Visualization methods for visual DNNs	3
Linardatos et al. (2021)	P, R	Technical, with source code	3
Bodria et al. (2021)	B, P	Comparative studies of XAI methods	4
Very long			
Molnar (2020)	B	Fundamental concepts of interpretability	5
Arrieta et al. (2020)	P, R	Notion of responsible AI: Terminology, broad collection of examples	1
Burkart and Huber (2021)	R	Supervised machine learning; linear, rule-based methods and decision trees, data analytics and ontologies	3
Vilone and Longo (2020)	R	Systematic and broad review of XAI surveys, theory, methods, and method evaluation	3

Sorted frst by length, then reception (citations per year). Detail is manually ranked in values from 1 (short discussion per method) to 5 (detailed discussion per method). The audience can be beginner (B), practitioner (P), or researcher (R)

Short surveys

A short overview mostly on explanations of supervised learning approaches was provided by Došilović et al. (2018). This quite short but broad and beginner-friendly introductory survey shortly covers diverse approaches and open challenges toward XAI. Slightly earlier in 2017, Biran and Cotton published their XAI survey (Biran and Cotton 2017). This is a short and early collection of explainability concepts. Their general focus lies on explanations of single predictions, and diverse model types like

rule-based systems and Bayesian networks, which are each shortly discussed. Most recently, in 2020 Benchekroun et al. collected and presented XAI methods from an industry point of view (Benchekroun et al. 2020). They present a preliminary taxonomy that includes pre-modelling explainability as an approach to link knowledge about data with knowledge about the used model and its results. Regarding the industry perspective, they specifcally motivate standardization.

Medium-length surveys

An important and very well received earlier work on XAI method collections was provided by Gilpin et al. (2018). Their extensive survey includes very different kinds of XAI methods, including, e.g., rule extraction. In addition, for researchers they provide references to more specialized surveys in the feld. It is very similar to the long survey by Adadi and Berrada (2018), only shortened by skipping detail on the methods. More detail, but a slightly more specialized focus, is provided in the survey by Du et al. (2019). The beginner-friendly high-level introduction to XAI features few, in detail discussed examples. These concentrate on perturbation based attention visualization. Similarly, the examples in Murdoch et al. (2019) also mostly focus on visual domains and explanations. This review by Murdoch et al. in 2019 is also beginner-friendly, and prefers detail over covering a high number of methods. The examples are embedded into a comprehensive short introductory review of key categories and directions in XAI. Their main focus is on the proposal of three simple practical desiderata for explanations: The model to explain should be predictive (predictive accuracy), the explanations should be faithful to the model (descriptive accuracy), and the information presented by the explanations should be relevant to the receiver. Slightly earlier, Goebel et al. (2018) concentrate in their work more on multimodal explanations and question-answering systems. This survey contains a high-level review of the need for XAI, and discussion of some exemplary state-of-the-art methods. A very recent and beginner-friendly XAI method review is by Islam et al. (2021). Their in-depth discussion of example methods is aligned with a shallow taxonomy, and many examples are practically demonstrated in a common simple case study. Additionally, they present a short meta-study of XAI surveys, and a collection of future perspectives for XAI. The latter includes formalization, XAI applications for fair and accountable ML, XAI for human-machine-systems, and more interdisciplinary cooperation in XAI research.

Long surveys

A well-received lengthy and extensive XAI literature survey was conducted by Adadi and Berrada (2018). They reviewed and shortly discuss 381 papers related to XAI to provide a holistic view on the XAI research landscape at that time. This includes methods, metrics, cognitive aspects, and aspects related to human-machine-systems. Another well-received, but more recent, slightly less extensive and more beginner-friendly survey is that by Carvalho et al. (2019). They collect and discuss in detail different aspects of XAI, especially motivation, properties, desirables, and metrics. Each aspect is accompanied by some examples. Similarly beginner-friendly is the introductory work of Xie et al. (2020). Their feld guide explicitly targets newcomers to the feld with a general introduction to XAI, and a wide variety of examples of standard methods, mostly for explanations of DNNs. A more formal introduction is provided by Das and Rad (2020) in their survey. They collect formal defnitions of XAI

related terms, and develop a shallow taxonomy. The focus of the examples is on visual local and global explanations of DNNs based on (model-specifc) backpropagation or (model-agnostic) perturbation-based methods. More recent and much broader is the survey of Linardatos et al. (2021). It provides an extensive technical collection and review of XAI methods with code and toolbox references, which makes it specifcally interesting for practitioners. Similarly and in the same year, Bodria et al. (2021) review in detail more than 60 XAI methods for visual, textual, and tabular data models. These are selected to be most recent and widely used, and cover a broad range of explanation types. Several comparative benchmarks of methods are included, as well as a short review of toolboxes.

Very long surveys

By now there are a couple of surveys available that aim to give a broad, rigorous, and in-depth introduction to XAI. A frst work in this direction is the regularly updated book on XAI by Molnar (2020), frst published in 2017.[5] This book is targeted at beginners, and gives a basic and detailed introduction on interpretability methods, including many transparent and many model-agnostic ones. The focus lies more on fundamental concepts of interpretability and detail on standard methods than on the amount of discussed methods. Meanwhile, Arrieta et al. (2020) put up a very long and broad collection of XAI methods in their 2020 review. The well-received survey can also be considered a base work of state-of-the-art XAI, as they introduce the notion of *responsible AI*, *i.e.*, development of AI models respecting fairness, model explainability, and accountability. This is also the focus of their work, in which they provide terminology, a broad but less detailed selection of example methods, and practical discussion for responsible AI. A very recent extensive XAI method survey is that of Burkart and Huber (2021). They review in moderate detail explainability methods, primarily for classifcation and regression in supervised machine learning. Specifcally, they include many rule-based and decision-tree based explanation methods, as well as aspects on data analysis and ontologies for formalizing input domains. This is preceded by a deep collection of many general XAI aspects. Finally and a little earlier, Vilone and Longo (2020) published in 2020 an equally extensive systematic literature survey on XAI in general. The systematic literature search was similar to ours, only with a different focus. They include a broad meta-study of reviews, as well as reviews and discussion of works on XAI theory, methods, and method evaluations.

4.3 Method collections with specific focus

Besides the many broad method collections, there by now are numerous ones specifically concentrating on an application domain, specifc input or task types, certain surrogate model types, or other traits of XAI methods. We here manually clustered surveys by similarity of their main focus points. An overview on the resulting clusters is given in Table 4.

[5] https://github.com/christophM/interpretable-ml-book/tree/v0.1.

Table 4 Overview on survey clusters reviewed Sect. 4.3 with restricted (restr.) focus

Restr. by:	Restr. to:	
Application domain	NLP	Danilevsky et al. (2020)
	Medicine	Singh et al. (2020), Tjoa and Guan (2020)
	Recommendation systems	Nunes and Jannach (2017), Zhang and Chen (2020)
Application type	Interactive ML	Anjomshoae et al. (2019), Baniecki and Biecek (2020), Amershi et al. (2014)
Task	Visual tasks	Samek et al. (2019), Samek and Müller (2019), Nguyen et al. (2019), Ancona et al. (2019), Alber (2019), Li et al. (2020), Zhang and Zhu (2018)
	Reinforcement ML	Puiutta and Veith (2020); Heuillet et al. (2021)
Explanator output type	Rule-based XAI	Cropper et al. (2020), Vassiliades et al. (2021/ed), Hailesilassie (2016), Calegari et al. (2020)
	Counterfactual explanations	Byrne (2019), Artelt and Hammer (2019), Verma et al. (2020), Keane et al. (2021), Mazzine and Martens (2021), Karimi et al. (2021), Stepin et al. (2021)
Other XAI method traits	Model-agnostic methods	Guidotti et al. (2018)

XAI for specific application domains

Some method surveys focus on concrete practical application domains. One is by Danilevsky et al. (2020), who survey XAI methods for *natural language processing* (NLP). This includes a taxonomy, a review of several metrics, and a dense collection of XAI methods. Another important application domain is the *medical domain*. For example, Singh et al. (2020) provided a survey and taxonomy of XAI methods for image classifcation with a focus on medical applications. A slightly broader focus on general XAI methods for medical applications was selected by Tjoa and Guan (2020) in the same year. Their long review shortly discusses more than sixty methods, and sorts them into a shallow taxonomy. Another domain sparking needs for XAI is that of *recommendation systems*, *e.g.*, in online shops. An example here is the long, detailed, and practically oriented survey by Nunes and Jannach (2017). Amongst others, they present a detailed taxonomy of XAI methods for recommendation systems (cf. Nunes and Jannach 2017, Fig. 11). A similar, but even more extensive and lengthy survey was provided by Zhang and Chen (2020). They generally review recommendation systems also in a practically oriented manner, and provide a good overview on models that are deemed explainable.

XAI for interactive ML applications

Several studies concentrate on XAI to realize interactive machine learning. For example, Anjomshoae et al. (2019) reviewed explanation generation, communication and evaluation for autonomous agents and human-robot interaction. Baniecki and Biecek (2020) instead directly concentrated on interactive machine learning in their medium-

length survey on the topic. They present challenges in explanation, traits to overcome these, as well as a taxonomy for interactive explanatory model analysis. The longer but earlier review by Amershi et al. (2014) more concentrates on practical case studies and research challenges. They motivate incremental, interactive and practical human-centered XAI methods.

XAI for visual tasks

Some of the earlier milestones for the current feld of XAI were methods to explain input importance for models with image inputs (Das and Rad 2020), such as LIME (Ribeiro et al. 2016) and LRP (Bach et al. 2015). Research on interpretability and explainability of models for visual tasks is still very active, as several surveys with this focus show. One collection of both methods and method surveys with a focus on visual explainability is the book edited by Samek et al. (2019). This includes the following surveys:

- Samek and Müller (2019): A short introductory survey on visual explainable AI for researchers, giving an overview on important developments;
- Nguyen et al. (2019): A survey specifcally on feature visualization methods. These are methods to fnd prototypical input patterns that most activate parts of a DNN. The survey includes a mathematical perspective on the topic and a practical overview on applications.
- Ancona et al. (2019): A detailed survey on gradient-based methods to fnd attribution of inputs to outputs; and
- Alber (2019): A detailed collection of implementation considerations regarding different methods for highlighting input attribution. The review includes code snippets for the TensorFlow deep learning framework.

Similar to Ancona et al. (2019), Li et al. (2020) focus on XAI methods to obtain heatmaps as visual explanations. They in detail discuss seven examples of methods, and conduct an experimental comparative study with respect to fv e specialized metrics. Another survey focusing on visual interpretability is the earlier work by Zhang and Zhu (2018). This well-received medium-length survey specializes on visual explanation methods for convolutional neural networks.

XAI for reinforcement learning tasks

Just as for visual tasks, there are some studies specifcally focusing on explanations in tasks solved by reinforcement learning. One is that by Puiutta and Veith (2020). This medium long to lengthy review provides a short taxonomy on XAI methods for reinforcement learning. It reviews more than 16 methods specifc to reinforcement learning in a beginner-friendly way. A comparable and more recent, but slightly longer, more extensive, and more technical survey on the topic is by Heuillet et al. (2021).

XAI methods based on rules

One type of explanation outputs is that of (formal) symbolic rules. Both generation of interpretable rule-based models, as well as extraction of approximate rule sets from less interpretable models have a long history. We here collect some more recent reviews on these topics. One is the historical review by Cropper et al. (2020) on the developments in Inductive Logic Programming. Inductive logic programming summarizes methods to automatically construct rule-sets for solving a task given some

formal background knowledge and few examples. The mentioned short survey aims to look back at the last 30 years of development in the feld, and serve as a good starting point for beginners. On the side of inherently interpretable rule-based models, Vassiliades et al. (2021/ed) recently reviewed argumentation frameworks in detail. An argumentation framework provides an interpretable logical argumentation line that formally deduces a statement from (potentially incomplete) logical background knowledge. This can, *e.g.*, be used to fnd the most promising statement from some choices, and, hence, as an explainable (potentially interactive) model on symbolic data. The long, detailed, and extensive survey formally introduces standard argumentation frameworks, reviews existing methods and applications, and promotes argumentation frameworks as promising interpretable models. While the previous surveys concentrate on directly training inherently interpretable models consisting of rules, Hailesilassie (2016) shortly reviewed rule extraction methods. Rule extraction aims to approximate a trained model with symbolic rule sets or decision trees. These methods have a long history but faced their limits when applied to large-sized models like state-of-the-art neural networks, as discussed in the survey. A more recent survey that covers both rule extraction as well as integration of symbolic knowledge into learning processes is the review by Calegari et al. (2020). Their long and in-depth overview covers the main symbolic/sub-symbolic integration techniques for XAI, including rule extraction methods for some steps.

Counterfactual and contrastive explanations

While the previous surveys were mostly concentrated on a specifc type of task, there are also some that are restricted to certain types of explainability methods. One rising category is that of contrastive and counterfactual explanations (also *counterfactuals*). The goal of these is to explain for an instance "why was the output P rather than Q?" (Stepin et al. 2021), and, in particular for counterfactual explanations, how input features can be changed to achieve Q (Mazzine and Martens 2021). To do this, one or several other instances are provided to the user that produce the desired output. One is the short survey by Byrne (2019). This reviews specifcally counterfactual explanations with respect to evidence from human reasoning. The focus here lies on additive and subtractive counterfactual scenarios. Also as early as 2019, Artelt and Hammer (2019) provide a beginner-friendly review on model-specifc counterfactual explanation methods. They consider a variety of standard ML models, and provide detailed mathematical background for each of them. A more broad survey on counterfactuals was conducted by Verma et al. (2020). The mid-length survey reviews 39 methods and discusses common desiderata, a taxonomy, evaluation criteria, and open challenges for this XAI subtopic. In particular, they suggest the following desiderata: counterfactual examples should be valid inputs that are similar to the training data; the example should be as similar as possible to the original (proximity), while changing as few features as possible (sparsity); feature changes should be actionable, *i.e.*, the explainee should be able to achieve them (*e.g.*, increase age, not decrease); and they should be *causal*, acknowledging known causal relationships in the model. Later studies confrm these desirables. In particular, Keane et al. (2021) review in total 100 methods with respect to common motivations for counterfactual explanations and typical shortcomings thereof, in order to guide researchers. In their short survey, they fnd that better

psychological grounding of counterfactuals as well as their evaluation is required, in specifc for validity and feature selection. Also, methods up to that point in time are often missing user studies and comparative tests. A closer look at how to tackle comparative studies was taken by Mazzine and Martens (2021) in their extensive survey. They benchmarked open source implementations of 10 strategies for counterfactual generation for DNNs on 22 different tabular datasets. The controlled experimental environment may serve as a role model for researchers for future evaluations. In contrast to functional aspects of counterfactual methods, Karimi et al. (2021) in parallel focused on the use-case perspective: The mid-length survey highlights the impact of the mentioned desiderata on the use-case of algorithmic recourse. Algorithmic recourse here means to provide explanations and actionable recommendations for individuals who encountered an unfavorable treatment by an automated decision-making system. Lastly, the broad and extensive survey by Stepin et al. (2021) unites research perspectives on conceptual and methodological work for advanced researchers. They rigorously survey 113 studies on counterfactual and contrastive explanations reaching back as far as the 1970s. The reader is provided with a review of terms and defnitions used throughout the primary literature, as well as a detailed taxonomy. The obtained conceptual insights are then related to the methodological approaches.

Model-agnostic XAI methods

A very well received, long and extensive survey for model-agnostic XAI methods on tabular data is by Guidotti et al. (2018). Besides the method review, they also develop a formal approach to defne XAI use-cases that is especially useful for practitioners.

4.4 Toolboxes

A single XAI method often does not do the full job of making all relevant model aspects clear to the explainee. Hence, such toolboxes have become usual that implement more than one explainability method in a single library with a common interface. For detailed lists of available toolboxes the reader is referred to, e.g., the related work in Arya et al. (2019, Tab. 1), the repository links in Linardatos et al. (2021, Tabs. A.1, A.2), and the toolbox review in Bodria et al. (2021, Sec. 7). For implementation considerations in the case of visual interpretability the review (Alber 2019) is a suitable read. We here provide some examples of publications presenting toolboxes published since 2019 that we analyzed for taxonomy aspects, summarized in Table 5.

Already in 2018, frst toolboxes like Skater (Choudhary 2018) were available for several XAI tasks. Skater, in specifc, provides in total seven XAI methods for both global and local explanation of different kinds of trained models, including DNNs for visual and textual inputs. The beginner-friendly Microsoft toolbox InterpretML (Nori et al. 2019) implements fv e model-agnostic and four transparent XAI methods that are shortly introduced in the paper. Another toolbox from that year is iNNvestigate by Alber et al. (2019). They specifcally concentrate on some standard post-hoc heatmapping methods for visual explanation. Arya et al. (2019) presented the IBM AI explainability 360 toolbox with 8 diverse XAI methods in 2019. The implemented workfo w follows a proposed practical, tree-like taxonomy of XAI methods. Implemented methods cover a broad range of explainability needs, including explainability

Table 5 Overview on the sub-selection of papers introducing explainability toolboxes that were considered in Sect. 4.4 with their respective code repository; more extensive overviews can be found, e.g., in (Arya et al. 2019, Tab. 1), (Linardatos et al. 2021, Tabs. A.1, A.2), and (Bodria et al. 2021, Sec. 7)

Toolbox	Publication	Code repository
Skater	Choudhary (2018)	https://github.com/oracle/Skater
InterpretML	Nori et al. (2019)	https://github.com/interpretml/interpret
iNNvestigate	Alber et al. (2019)	https://github.com/albermax/innvestigate
AI Fairness 360	Arya et al. (2019)	https://github.com/Trusted-AI/AIF360
explAIner	Spinner et al. (2020)	https://github.com/dbvis-ukon/explainer
FAT Forensics	Sokol et al. (2020)	https://github.com/fat-forensics/fat-forensics
Alibi	Klaise et al. (2021)	https://github.com/SeldonIO/alibi

of data, inherently interpretable models, and post-hoc explainability both globally and locally. More recently, Spinner et al. (2020) presented their toolbox explAIner. This is realized as a plugin to the existing TensorBoard[6] toolkit for the TensorFlow deep learning framework. The many post-hoc DNN explanation and analytics tools are aligned with the suggested pipeline phases of model understanding, diagnosis, and refnement. Target users are both researchers, practitioners, and beginners. In 2020 also the FAT Forensics (for Fairness, Accountability, Transparency) toolbox was published (Sokol et al. 2020). Based on scikit-learn, several applications toward model and data quality checks and local and global explainability are implemented. The focus is on black-box methods, providing a generic interface for two use-cases: research on new fairness metrics (for researchers) and monitoring of ML models during preproduction (for practitioners). At the time of writing, methods for image and tabular data are supported. The most recent toolbox considered here is the Alibi explainability kit by Klaise et al. (2021). It features nine diverse, mostly black-box XAI methods for classifcation and regression. This includes both local and (for tabular data) global state-of-the-art analysis methods for image or tabular input data, with some additionally supporting textual inputs. Alibi targets practitioners, aiming to be an extensively tested, production-ready, scalable, and easy to integrate toolbox for explanation of machine learning models.

5 Taxonomy

This section establishes a taxonomy of XAI methods that unites terms and notions from the literature and that helps to differentiate and evaluate XAI methods. More than 70 surveys that have been selected in the course of our literature search (cf. Sect. 3), including the ones discussed in Sect. 4, were analyzed for such terms. We then identifed synonymous terms. Finally, we sub-structured the notions systematically according to practical considerations in order to provide a complete picture of the state-of-the-art XAI method and evaluation aspects. This section details the found

[6] TensorBoard toolkit: https://www.tensorfo w.org/tensorboard.

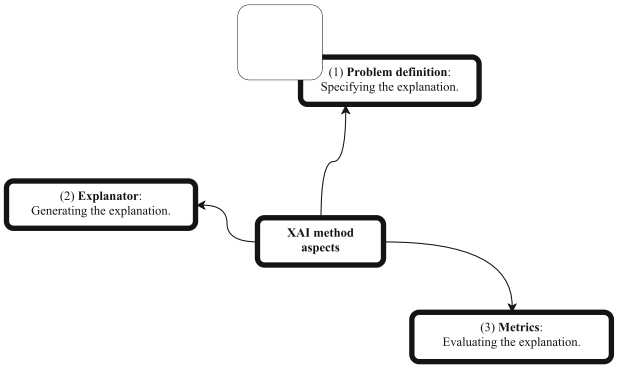

Fig. 3 Overview of the top-level categorization of taxonomy aspects explained in Sect. 5. Find a visualization of the complete taxonomy in Fig. 7

notions and synonyms, and our proposed structure thereof, which defnes the outline. An overview of the root-level structure is provided in Fig. 3, and a complete overview of the fnal structure in Fig. 7. The sub-trees for the frst-le vel categories can be found at the beginning of each section.

At the root level, we propose to categorize in a *procedural manner* according to the steps for building an explanation system. The following gives a high-level overview of the frst two levels of the taxonomy structure and the outline of this chapter (cf. overview in Fig. 3):

1. **Problem definition** (Sect. 5.1): One usually should start with the *problem definition*. This encompasses
   - traits of the *task*, and
   - the *explanandum* (precisely: the *interpretability* of the explanandum).
2. **Explanator properties** (Sect. 5.2): Then, we detail the *explanator properties* (Sect. 5.2), which we functionally divided into properties of
   - *input*,
   - *output*,
   - *interactivity* with the user, and
   - any further *formal constraints* posed on the explanator.
3. **Metrics** (Sect. 5.3): Lastly, we discuss different *metrics* (Sect. 5.3) that can be applied to explanation systems in order to evaluate their qualities. Following Doshi-Velez and Kim (2017), these are divided by their dependence on subjective human evaluation and the application into:
   - *functionally-grounded* metrics (independent of human judgment),
   - *human-grounded* metrics (subjective judgment required), and
   - *application-grounded* (full human-AI-system required).

The presented aspects are illustrated by selected example methods. The selection of example methods is by no means complete. Rather it intends to give an impression

about the wide range of the topic and how to apply our taxonomy to both some well-known and less known but interesting methods. An overview of the discussed method examples is given in Table 6.

On requirements derivation
Note that from a procedural perspective, the frst step when designing an explanation system should be to determine the use-case-specifc *requirements* (as part of the problem defnition). The requirements can be derived from all the taxonomy aspects that are collected in this review. This gives rise to a similar sub-structuring as the procedural one shown above:

- Both explanandum and explanator must match the *task*,
- the explanandum should be chosen to match the (inherent) *interpretability* needs, and
- the explanator must fulfll any other (functional and architectural) *explanator constraints* (see Sect. 5.2), as well as
- any *metric target values*.

This should be motivated by the actual goal or desiderata of the explanation. These can be, *e.g.*, verifability of properties like fairness, safety, and security, knowledge discovery, promotion of user adoption respectively trust, or many more. An extensive list of desiderata can be found in the work of Langer et al. (2021). As noted in Sect. 2.1, a detailed collection of XAI needs, desiderata, and typical use-cases is out of the scope of this work. Instead, our collection of taxonomy aspects shall serve as a starting point for effective and complete use-case analysis and requirements derivation.

5.1 Problem definition

The following aspects consider the concretion of the explainability problem (Fig. 4). Apart from the use-case analysis, which is skipped in this work, details on the following two aspects must be clear:

- **Task**: the *task* that is to be explained must be clear, and
- **Model interpretability**: the solution used for the task, meaning the *type of explanandum*. For explainability purposes, the level of *interpretability* of the explanandum model is the relevant point here.

5.1.1 Task

Out-of-the-box, XAI methods usually only apply to a specifc set of

- *task types* of the to-be-explained model, and
- *input data types*.

For white-box methods that access model internals, additional constraints may hold for the *architecture* of the model [cf. portability aspect in Yao (2005)].

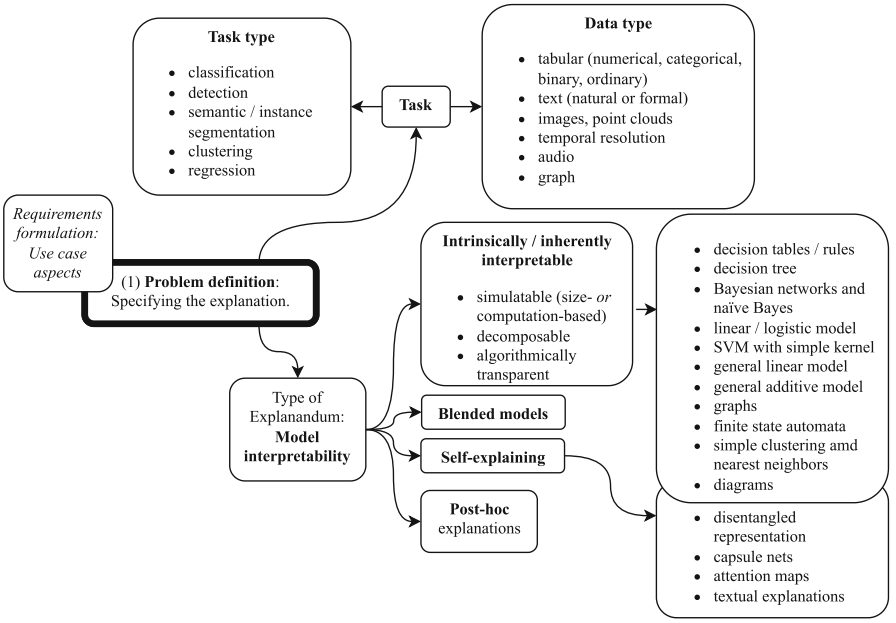

Fig. 4 Overview of the taxonomy aspects related to the problem defnition that are detailed in Sect. 5.1. Find a visualization of the complete taxonomy in Fig. 7

Task type

Typical task categories are unsupervised clustering (clu), regression, classifcation (cls), detection (det), segmentation (seg), either semantic, which is pixel-wise classifcation, or segmentation of instances. Many XAI methods that target a question for classifcation, *e.g.*, "Why this class?" can be extended to det, seg, and temporal resolution. This can be achieved by snippeting of the new dimensions: "Why this class in this spatial/temporal snippet?". It must be noted that XAI methods working on classifers often require access to the prediction of a continuous classifcation score instead of the fnal discrete classifcation. Such methods can also be used on regression tasks to answer questions about local trends, *i.e.*, "Why does the prediction tend in this direction?". Examples of regression predictions are bounding box dimensions in object detection.

Examples

RISE (Petsiuk et al. 2018). RISE (Randomized Input Sampling for Explanation) by Petsiuk et al. (2018) is a model-agnostic attribution analysis method specializing in image classifcation. For an input image, it produces a heatmap that highlights those superpixels in the image, which, when deleted, have the greatest infuence on the class confdence. High-attribution superpixels are found by randomly dimming superpixels of the input image.

D-RISE (Petsiuk et al. 2021). The method D-RISE by Petsiuk et al. (2021) extends this to object detection. It considers not a one-dimensional class confdence but the total prediction vector of a detection. The infuence of dimming is measured as the distance between the prediction vectors.

ILP (Cropper et al. 2020). The mentioned image-specifc (*i.e.*, local) explanation methods use the continuous class or prediction scores of the explanandum, and, hence, are in principle also applicable to regressors. In contrast, surrogate models produced using inductive logic programming (ILP) (Cropper et al. 2020) require the binary classifcation output of a model. ILP frameworks require input background knowledge (logical theory), together with positive and negative examples. From this, a logic program in the form of frst-order rules is learned that covers as many of the positive samples as possible.
CA-ILP (Rabold et al. 2020). An example of an ILP-based XAI method for convolutional image classifer is CA-ILP (Concept Analysis for ILP) by Rabold et al. (2020). In order to explain parts of the classifer with logical rules, they frst train small global models that extract symbolic features from the DNN intermediate outputs. These feature outputs are then used to train an ILP surrogate model. Lastly, clustering tasks can often be explained by providing examples or prototypes of the fnal clusters, which will be discussed in Sect. 5.2.

Input data type
Not every XAI method supports every input and output *signal type*, also called data type (Guidotti et al. 2018). One input type is tabular (symbolic) data, which encompasses numerical, categorical, binary, and ordinary (ordered) data. Other symbolic input types are natural language or graphs, and non-symbolic types are images and point clouds (with or without temporal resolution), as well as audio.

Examples Typical examples for image explanations are methods producing heatmaps. These highlight parts of the image that were relevant to the decision or a part thereof. This highlighting of input snippets can also be applied to textual inputs where single words or sentence parts may serve as snippets.
LIME (Ribeiro et al. 2016). A prominent example of heatmapping that is both applicable to images and text inputs is the model-agnostic LIME (Ribeiro et al. 2016) method (Local Interpretable Model-agnostic Explanations). It locally approximates the explanandum model by a linear model on feature snippets of the input. For training of that linear model, randomly selected snippets are removed. In the case of textual inputs, the words are considered as snippets, and for images pixels are grouped into superpixels. The removal of superpixels is here realized by coloring them with a neutral color, e.g., black. While LIME is suitable for image or textual input data, Guidotti et al. (2018) provide a broad overview of model-agnostic XAI methods for tabular data.

5.1.2 Model interpretability

Model interpretability here refers to the level of interpretability of the explanandum, *i.e.*, the model used to solve the original task of the system. A model is interpretable if it gives rise not only to mechanistic understanding (transparency) but also to a functional understanding by a human (Páez 2019). Explainability of (aspects of) the explanation system can be achieved by one of the following choices:

- start from the beginning with an *intrinsically interpretable* explanandum model (also called *ante-hoc interpretable* (Burkart and Huber 2021) or *intrinsically interpretable*) or a
- *blended*, *i.e.*, partly interpretable model (also called *interpretable by design* (Burkart and Huber 2021));
- design the explanandum model to include *self-explanations* as additional output; or
- *post-hoc* fnd an interpretable helper model without changing the trained explanandum model.

Intrinsic or inherent interpretability

As introduced by Lipton (2018), one can further differentiate between different levels of model transparency. The model can be understood as a whole, *i.e.*, a human can adopt it as a mental model [*simulatable* (Arrieta et al. 2020)]. Alternatively, it can be split up into simulatable parts [*decomposable* (Arrieta et al. 2020)]. Simulatability can either be measured based on the size of the model or based on the needed length of computation (cf. discussion of metrics in Sect. 5.3). As a third category, *algorithmic transparency* is considered, which means the model is mathematically understood, *e.g.*, the shape of the error surface is known. This is considered the weakest form of transparency, because the algorithm may not be simulatable as a mental model. The following models are considered inherently transparent in the literature [cf. Molnar (2020, Chap. 4), Guidotti et al. (2018, Sec. 5), Nori et al. (2019)]:

Decision tables and rules as experimentally evaluated by Huysmans et al. (2011), Allahyari and Lavesson (2011), Freitas (2014); This encompasses boolean rules as can be extracted from decision trees or fuzzy or frst-order logic rules. For further insights into inductive logic programming approaches to fnd the latter kind of rules, see, *e.g.*, the recent survey by Cropper et al. (2020).

Decision trees as empirically evaluated by Huysmans et al. (2011), Freitas (2014), Allahyari and Lavesson (2011);

Bayesian networks and naïve Bayes models as of Burkart and Huber (2021); interpretability of Bayesian network classifers was, *e.g.*, experimentally evaluated by Freitas (2014).

Linear and logistic models as of, *e.g.*, Molnar (2020);

Support vector machines as of Singh et al. (2020); as long as the used kernel function is not too complex, non-linear SVMs give interesting insights into the decision boundary.

General linear models (GLM) to inherently provide weights for the importance of input features; Here, it is assumed that there is a transformation, such that there is a linear relationship between the transformed input features and the expected output value. For example, in logistic regression, the transformation is the logit. See, *e.g.*, Molnar (2020, Sec. 4.3) for a basic introduction and further references.

General additive models (GAM) also inherently come with feature importance weights; Here, it is assumed that the expected output value is the sum of transformed features. See the survey by Chang et al. (2020) for more details and further references.

Examples
Additive Model Explainer (Chen et al. 2019b). One concrete example of general additive models is the Additive Model Explainer by Chen et al. (2019b). They train predictors for a given set of features, and another small DNN predicts the additive weights for the feature predictors. They use this setup to learn a GAM surrogate model for a DNN, which also provides a prior to the weights: They should correspond to the sensitivity of the DNN with respect to the features.

Graphs as of, *e.g.*, Xie et al. (2020);
Finite state automata as of Wang et al. (2018b);
Simple clustering and nearest neighbors approaches as of Burkart and Huber (2021);

Examples Examples are k-nearest neighbors (supervised) or k-means clustering (unsupervised).
k-means clustering (Hartigan and Wong 1979). The standard k-means clustering method introduced by Hartigan and Wong (1979) works with an intuitive model, simply consisting of k prototypes and a proximity measure, with inference associating new samples to the closest prototype representing a cluster.
k-NN (Altman 1992) k-nearest neighbors (k-NN) determines for a new input the k samples from a labeled database that are most similar to the new input sample. The majority vote of the nearest labels is then used to assign a label to the new instance. As long as the proximity measure is not too complex, these methods can be regarded as unsupervised respectively supervised inherently interpretable models. k-NN was experimentally evaluated for interpretability by Freitas (2014).

Diagrams as of Heuillet et al. (2021).

Blended models
Blended models (also called *interpretable by design* (Burkart and Huber 2021)) consist partly of intrinsically transparent, symbolic models that are integrated in sub-symbolic non-transparent ones. These kinds of hybrid models are especially interesting for neuro-symbolic computing and similar felds combining symbolic with sub-symbolic models (Calegari et al. 2020).

Examples
Logic Tensor Nets (Donadello et al. 2017). An example of a blended model is the Logic Tensor Network. Their idea is to use fuzzy logic to encode logical constraints on DNN outputs, with a DNN acting as a fuzzy logic predicate. The framework by Donadello et al. (2017) allows additionally to learn semantic relations subject to symbolic fuzzy logic constraints. The relations are represented by simple linear models.
FoldingNet (Yang et al. 2017), **Neuralized clustering** (Kauffmann et al. 2019). Unsupervised deep learning can be made interpretable by several approaches, *e.g.*, combining autoencoders with visualization approaches. Another approach explains choices of "neuralized" clustering methods (Kauffmann et al. 2019) (*i.e.*, clustering models translated to a DNN) with saliency maps. Enhancing

an autoencoder was applied, for example, in the FoldingNet (Yang et al. 2017) architecture on point clouds. There, a folding-based decoder allows for viewing the reconstruction of point clouds, namely the warping from a 2D grid into the point cloud surface. A saliency-based solution can be produced by algorithms such as layer-wise relevance propagation, which will be discussed in later examples.

Self-explaining models

Self-explaining models provide additional outputs that explain the output of a single prediction. According to Gilpin et al. (2018), there are three standard types of outputs of explanation generating models: *attention maps*, *disentangled representations*, and *textual or multi-modal explanations*.

Attention maps These are heatmaps that highlight the relevant parts of a given single input for the respective output.

Examples The work by Kim and Canny (2017) adds an attention module to a DNN that is processed in parallel to, and later multiplied with, convolutional outputs. Furthermore, they suggest a clustering-based post-processing of the attention maps to highlight the most meaningful parts.

Disentangled representations Representations in the intermediate output of the explanandum are called disentangled if single or groups of dimensions therein directly represent symbolic (also called semantic) concepts.

Examples One can, by design, force one layer of a DNN to exhibit a disentangled representation.

Capsule Nets (Sabour et al. 2017). One example is the capsule network by Sabour et al. (2017), where groups of neurons, the capsules, characterize each an individual entity, *e.g.*, an object or object part. The length of a capsule vector is interpreted as the probability that the corresponding object is present, while the rotation encodes the properties of the object (*e.g.*, rotation or color). Later capsules get as input the weighted sum of transformed previous capsule outputs, with the transformations learned and the weights obtained in an iterative routing process. A simpler disentanglement than an alignment of semantic concepts with groups of neurons is the alignment of single dimensions.

ReNN (Wang 2018). This is done, *e.g.*, in the ReNN architecture developed by Wang (2018). They explicitly modularize their DNN to ensure semantically meaningful intermediate outputs.

Semantic Bottlenecks (Losch et al. 2019). Other methods rather follow a post-hoc approach that fne-tunes a trained DNN toward more disentangled representations, as suggested for Semantic Bottleneck Networks (Losch et al. 2019). These consist of the pretrained backbone of a DNN, proceeded by a layer in which each dimension corresponds to a semantic concept, called semantic bottleneck, and fnalized by a newly trained front DNN part. During fne-tuning, frst, the connections from the backend to the semantic bottleneck are trained, then the parameters of the front DNN.

Concept Whitening (Chen et al. 2020). Another interesting fne-tuning approach is that of concept whitening by Chen et al. (2020), which supple-

ments batch-normalization layers with a linear transformation that learns to align semantic concepts with unit vectors of an activation space.

Textual or multi-model explanations These provide the explainee with a direct verbal or combined explanation as part of the model output.

Examples (Kim et al. 2018b). An example are the explanations provided by Kim et al. (2018b) for the application of end-to-end steering control in autonomous driving. Their approach is two-fold: They add a custom layer that produces attention heatmaps similar to those from (Kim and Canny 2017); a second custom part uses these heatmaps to generate textual explanations of the decision, which are (weakly) aligned with the model processing.

ProtoPNet (Chen et al. 2019a). ProtoPNet by Chen et al. (2019a) for image classifcation provides visual examples rather than text. The network architecture is based on frst selecting prototypical image patches and then inserting a prototype layer that predicts similarity scores for patches of an instance with prototypes. These can then be used for explanation of the fnal result in the manner of "This is a sparrow as its beak looks like that of other sparrow examples".

(Hendricks et al. 2016). A truly multi-modal example is that by Hendricks et al. (2016), which trains alongside a classifer a long-short term memory DNN (LSTM) to generate natural language justifcations of the classifcation. The LSTM uses both the intermediate features and predictions of the image classifer and is trained toward high-class discriminativeness of the justifcations. The explanations can optionally encompass bounding boxes for features that were important for the classifcation decision, making it multi-modal.

Post-hoc

Post-hoc methods use a (local or global) helper model from which to derive an explanation. This explainable helper model can either aim to

- fully mimic the behavior of the explanandum, or
- only approximate sub-aspects like input attribution.

Helper models that fully approximate the explanandum are often called *surrogate* or *proxy* models, and the process of training them is termed *model distillation*, *student-teacher* approach, or *model induction*. However, it is often hard to differentiate between the two types. Hence, for consistency, we here use the terms proxy and surrogate model for any type of helper model. Many examples of post-hoc methods are given in the course of the upcoming taxonomy aspects.

5.2 Explanator

One can consider an explanator simply as an implemented function that outputs explanations. This allows to structure aspects of the explanator into the main defning aspects of a function:

- the *input*,

- the *output*,
- the function class described by *mathematical properties or constraints*, and
- the actual processing of the explanation function, like its *interactivity*.

An overview of the aspects discussed in this section is given in Fig. 5.

5.2.1 Input

The following explanator characteristics are related to the explanator input:

- What are the *required inputs* (the explanandum model, data samples, or even user feedback)?
- In how far is the method *portable* to other input types, *e.g.*, explanandum model types?
- In how far are explanations *local to an input instance or global for the complete input*?

Required input
The necessary inputs to the explanator may differ amongst methods (Spinner et al. 2020). The explanandum must usually be provided to the explanator. Many methods do also require valid *data* samples. Some even require *user feedback* (cf. Sect. 5.2.3) or further situational *context* (cf. (Dey 2001) for a more detailed defnition of context).

Portability
An important practical aspect of post-hoc explanations is whether or how far the explanation method is dependent on access to the internals of the explanandum model. This level of dependency is called portability, translucency, or transferability. In the following, we will not further differentiate between the strictness of requirements of model-specifc methods. Transparent and self-explaining models are always model-specifc, as the interpretability requires a special model type or model architecture (modifcation). Higher levels of dependency are:

Model-agnostic also called *pedagogical* (Yao 2005) or black-box: This means that only access to model input and output is required.

> *Examples* A prominent example of model-agnostic methods is the previously discussed LIME (Ribeiro et al. 2016) method for local approximation via a linear model.
> **SHAP** (Lundberg and Lee 2017). Another method to fnd feature importance weights without any access to model internals is SHAP (SHapley Additive exPlanation) by Lundberg and Lee (2017). Their idea is to axiomatically ensure: local fdelity; features missing from the original input have no effect; an increase in weight also means an increased attribution of the feature to the fnal output and uniqueness of the weights. Just as LIME, SHAP just requires a defnition of "feature" or snippet on the input in order to be applicable.

Model-specific also called *decompositional* (Yao 2005) or white-box: This means that access is needed to the internal processing or architecture of the explanandum model, or even constraints apply.

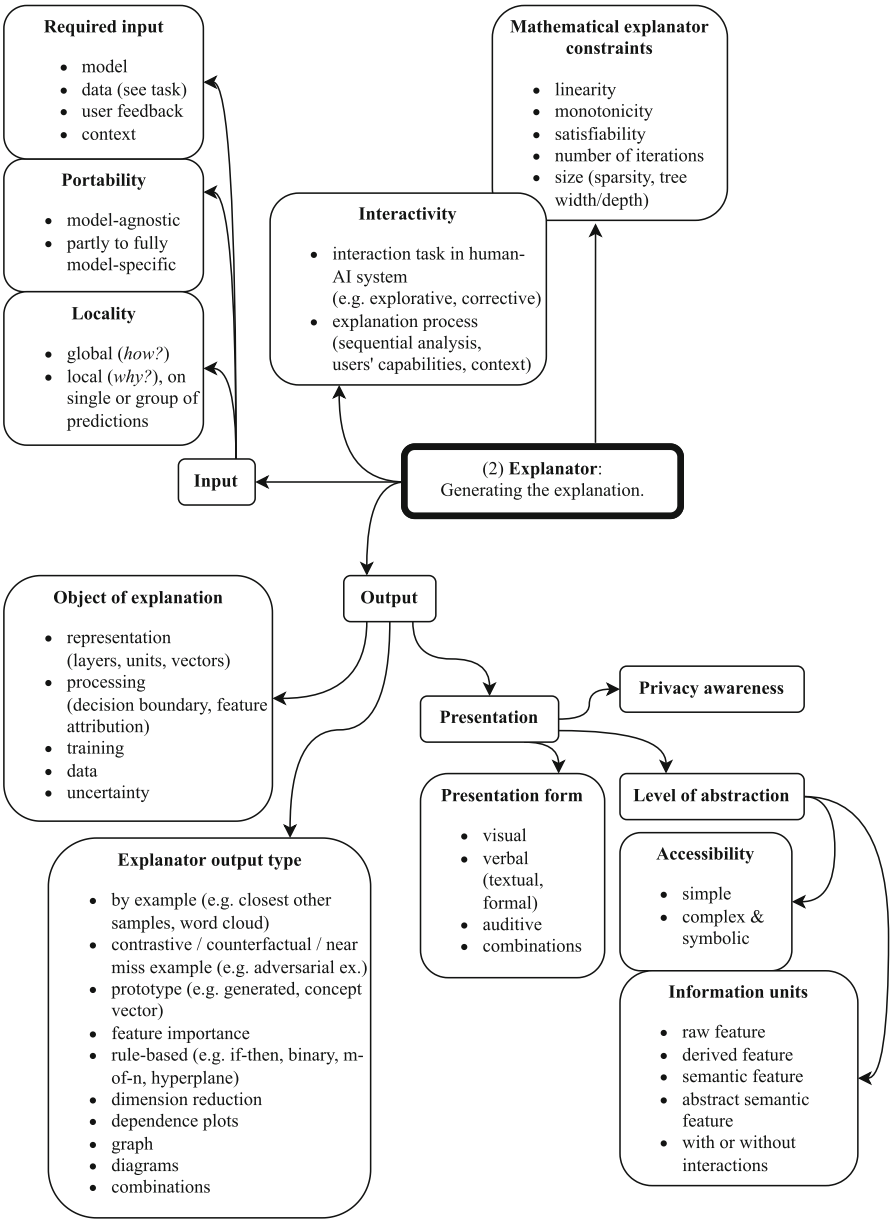

Fig. 5 Overview of the taxonomy aspects related to the explanator that are detailed in Sect. 5.3. Find a visualization of the complete taxonomy in Fig. 7

Examples Methods relying on gradient or relevance information for the generation of visual attention maps are strictly model-specifc.

Sensitivity Analysis (Baehrens et al. 2010). A gradient-based method is Sensitivity Analysis by Baehrens et al. (2010). They pick the vector representing the steepest ascension in the gradient tangential plane of a sample point. This method is independent of the type of input features but can only analyze a single one-dimensional output at once.

Deconvnet (Zeiler and Fergus 2014), **Backprop** (Simonyan et al. 2014), **Guided Backprop** (Springenberg et al. 2015). Deconvnet by Zeiler and Fergus (2014) instead is agnostic to the type of output, but depends on a convolutional architecture and image inputs. The same holds for its successors Backpropagation (Simonyan et al. 2014) and Guided Backpropagation (Springenberg et al. 2015). They approximate a reconstruction of input by defning inverses of pool and convolution operations. This allows for backpropagating the activation of single flters back to input image pixels (see (Weitz 2018) for a good overview).

LRP (Bach et al. 2015). The idea of Backpropagation is generalized axiomatically by LRP (Layer-wise Relevance Propagation): Bach et al. (2015) require that the sum of linear relevance weights for each neuron in a layer should be constant throughout the layers. The rationale behind this is that relevance is neither created nor extinguished from layer to layer. Methods that achieve this are, *e.g.*, Taylor decomposition or the backpropagation of relevance weighted by the forward-pass weights.

PatternAttribution (Kindermans et al. 2018). The advancement PatternAttribution by Kindermans et al. (2018) fulflls the additional constraint to be sound on linear models.

Hybrid also called *eclectic* (Yao 2005) or *gray-box*: This means the explanator only depends on access to parts of the model intermediate output, but not the full architecture.

Examples **DeepRED** (Zilke et al. 2016). The rule extraction technique DeepRED (Deep Rule Extraction with Decision tree induction) by Zilke et al. (2016) is an example of an eclectic method, so neither fully model-agnostic nor totally reliant on access to model internals. The approach conducts a backward induction over the layer outputs of a DNN, between each two applying a decision tree extraction. While they enable rule extraction for arbitrarily deep DNNs, only small networks will result in rules of decent length for explanations.

Explanation locality

Literature differentiates between different ranges of validity of an explanation respectively surrogate model. A surrogate model is valid in the ranges where high fdelity can be expected (see Sect. 5.3). The range of input required by the explanator depends on the targeted validity range, so whether the input must represent a *local* or the *global* behavior of the explanandum. The general locality types are:

Local An explanation is considered local if the explanator is valid in a neighborhood of one or a group of given (valid) input samples. Local explanations tackle the question of *why* a given decision for one or a group of examples was made.

Examples Heatmapping methods are typical examples for local-only explanators, such as the discussed perturbation-based model-agnostic methods RISE (Petsiuk et al. 2018), D-RISE (Petsiuk et al. 2021), LIME (Ribeiro et al. 2016), SHAP (Lundberg and Lee 2017), as well as the model-specifc sensitivity and backpropagation based methods LRP (Bach et al. 2015), PatternAttribution (Kindermans et al. 2018), Sensitivity Analysis (Baehrens et al. 2010), and Deconvnet and its successors (Zeiler and Fergus 2014; Simonyan et al. 2014; Springenberg et al. 2015).

Global An explanation is considered global if the explanator is valid in the complete (valid) input space. Other than the *why* of local explanations, global interpretability can also be described as answering *how* a decision is made.

Examples **Explanatory Graphs** (Zhang et al. 2018). A graph-based global explanator is generated by Zhang et al. (2018). Their idea is that semantic concepts in an image usually consist of sub-objects to which they have a constant relative spatial relation (*e.g.*, a face has a nose in the middle and two eyes next to each other) and that the localization of concepts should not only rely on high flter activation patterns, but also on their sub-part arrangement. To achieve this, they translate the convolutional layers of a DNN into a tree of nodes (concepts), the *explanatory graph*. Each node belongs to one flter , is anchored at a fx ed spatial position in the image, and represents a spatial arrangement of its child nodes. The graph can also be used for local explanations via heatmaps. For localizing a node in an input image, the node is assigned the position closest to its anchor for which (1) its flter activation is highest, and (2) the expected spatial relation to its children is best fulflled.
Feature Visualization (Olah et al. 2017). While most visualization-based methods provide only local visualizations, Feature Visualizations as reviewed by Olah et al. (2017) give a global, prototype-based, visual explanation. The goal here is to visualize the functionality of a DNN's part. It is achieved by fnding prototypical input examples that strongly activate that part. These can be found via picking, search, or optimization.
VIA (Thrun 1995). Other than visualizations, rule extraction methods usually only provide global approximations. An example is the well-known model-agnostic rule extractor VIA (Validity Interval Analysis) by Thrun (1995), which iteratively refnes or generalizes pairs of input- and output-intervals.
SpRAy (Lapuschkin et al. 2019). An example of getting from local to global explanations is SpRAy (Spectral Relevance Analysis) by Lapuschkin et al. (2019). They suggest to apply spectral clustering (von Luxburg 2007) to local feature attribution heatmaps of data samples in order to fnd spuriously distinct global behavioral patterns. The heatmaps were generated via LRP (Bach et al. 2015).

5.2.2 Output

The output is characterized by several aspects:

- what is explained (the *object of explanation*),
- how it is explained (the actual *output type*, also called *explanation content type*), and
- how it is *presented*.

Object of explanation

The object (or scope (Molnar 2020)) of an explanation describes which item of the development process should be explained. Items we identifed in the literature:

Processing The objective is to understand the (symbolic) processing pipeline of the model, *i.e.*, to answer parts of the question "How does the model work?". This is the usual case for model-agnostic analysis methods. Types of processing to describe are, *e.g.*, the *decision boundary* and *feature attribution* (or feature importance). Note that these are closely related, as highly important features usually locally point out the direction to the decision boundary. In case a symbolic explanator is targeted, one may need to frst fnd a symbolic representation of the input, output, or the model's internal representation. Note that model-agnostic methods that do not investigate the input data usually target explanations of the model processing.

Examples Feature attribution methods encompass all the discussed attribution heatmapping methods (*e.g.*, RISE (Petsiuk et al. 2018), LIME (Ribeiro et al. 2016), LRP (Bach et al. 2015)). LIME can be considered a corner case: In addition to explaining feature importance it approximates the decision boundary using a linear model on superpixels. The linear model itself may already serve as an explanation. Typical ways to describe decision boundaries are decision trees or sets of rules, as extracted by the discussed VIA (Thrun 1995), and DeepRED (Zilke et al. 2016) approaches.

TREPAN (Craven and Shavlik 1995). Standard candidates for model-agnostic decision tree extraction are TREPAN by Craven and Shavlik (1995), and C4.5 by Quinlan (1993). TREPAN uses M-of-N rules at the split points of the extracted decision tree.

C4.5 (Quinlan 1993). C4.5 uses interval-based splitting points, and generates shallower but wider trees compared to TREPAN.

Concept Tree (Renard et al. 2019). Concept tree is a recent extension of TREPAN by Renard et al. (2019) that adds automatic grouping of correlated features into the candidate concepts to use for the tree nodes.

Inner representation Machine learning models learn new representations of the input space, like the latent space representations found by DNNs. Explaining these inner representations answers, "How does the model see the world?". A more fne-grained differentiation considers whether *layers*, *units*, or *vectors* in the feature space are explained.

Examples

- Units: One example of unit analysis is the discussed Feature Visualization (Olah et al. 2017).

NetDissect (Bau et al. 2017). In contrast to this unsupervised assignment of convolutional filters to prototypes, NetDissect (Network Dissection) by Bau et al. (2017) assigns filters to pre-defned semantic concepts in a supervised manner: For a filter, that semantic concept (color, texture, material, object, or object part) is selected for which the ground truth segmentation masks have the highest overlap with the upsampled filter's activations. The authors also suggest that concepts that are less entangled, so less distributed over filters, are more interpretable, which is measurable with their filter-to-concept-alignment technique.

- Vectors:
  Net2Vec (Fong and Vedaldi 2018). Other than NetDissect, Net2Vec by Fong and Vedaldi (2018) also wants to assign concepts to their possibly entangled representations in the latent space. Given a concept, they train a linear 1×1-convolution on the output of a layer to segments the respective concept with an image. The weight vector of the linear model for a concept can be understood as a prototypical representation (embedding) for that concept in the DNN intermediate output. They found that such embeddings behave like vectors in a word vector space: Concepts that are semantically similar feature embeddings with high cosine similarity.
  TCAV (Kim et al. 2018a). Similar to Net2Vec, TCAV (Testing Concept Activation Vectors) also aims to fnd embeddings of NetDissect concepts. Kim et al. (2018a) are not interested in embeddings that are represented as a linear combination of convolutional filters, but instead in embedding vectors lying in the space of the complete layer output. In other words, they do not segment concepts, but make an image-level classifcation of whether the concept is present. These are found by using an SVM model instead of the 1×1-convolution. Additionally, they suggest using partial derivatives along those concept vectors to fnd the local attribution of a semantic concept to a certain output.
  ACE (Ghorbani et al. 2019). Other than the already mentioned supervised methods, ACE (Automatic Concept-based Explanations) by Ghorbani et al. (2019) does not learn a linear classifer but does an unsupervised clustering of concept candidates in the latent space. The cluster center is then selected as the embedding vector. A superpixeling approach is used together with outlier removal to obtain concept candidates.
- Layers:
  Concept completeness (Yeh et al. 2020), **IIN** (Esser et al. 2020). The works of Yeh et al. (2020) and the IIN (invertible interpretation networks) approach by Esser et al. (2020) extend on the previous approaches and analyze a complete layer output space at once. For this, they fnd a subspace with a basis of concept embeddings, which allows an invertible transformation to a disentangled representation space. While IIN uses invertible DNNs for the bijection of concept space to latent space, Yeh et al. (2020) use linear maps in their experiments. These approaches can be seen as a post-hoc version of the Semantic Bottleneck (Losch et al. 2019) architecture, only not replacing the complete later part of the model, but just

learning connections from the bottleneck to the succeeding trained layer. Yeh et al. (2020) additionally introduce the notion of completeness of a set of concepts as the maximum performance of the model intercepted by the semantic bottleneck.

Development (during training) Some methods focus on assessing the effects during training (Molnar 2020, Sec. 2.3): "How does the model evolve during the training? What effects do new samples have?"

Examples (Shwartz-Ziv and Tishby 2017). One example is the work of Shwartz-Ziv and Tishby (2017), who inspect the model during training to investigate the role of depth in neural networks. Their fndings indicate that depth actually is of computational beneft.

Influence Functions (Koh and Liang 2017). An example which can be used to provide, *e.g.*, prototypical explanations are Infuence Functions by Koh and Liang (2017). They gather the infuence of training samples during the training to later assess the total impact of samples on the training. They also suggest using this information as a proxy to estimate the infuence of the samples on model decisions.

Uncertainty Molnar (2020) suggests to capture and explain (*e.g.*, visualize) the uncertainty of a prediction of the model. This encompasses the broad feld of Bayesian deep learning (Kendall and Gal 2017) and uncertainty estimation (Henne et al. 2020). Several works argue why it is important to make the uncertainty of model decisions accessible to users. For example, Pocevičiūtė et al. (2020) argues this for medical applications, and McAllister et al. (2017) for autonomous driving.

Data Pre-model interpretability (Carvalho et al. 2019) is the point where explainability touches the large research area of data analysis and feature mining.

Examples **PCA** (Jolliffe 2002). Typical examples for projecting high-dimensional data into easy-to-visualize 2D space are component analysis methods like PCA (Principal Component Analysis) which was introduced by Jolliffe (2002).
t-SNE (van der Maaten and Hinton 2008). A slightly more sophisticated approach is t-SNE (t-Distributed Stochastic Neighbor Embedding) by van der Maaten and Hinton (2008). In order to visualize a set of high-dimensional data points, they try to fnd a map from these points into a 2D or 3D space that is faithful to pairwise similarities.
Spectral Clustering (von Luxburg 2007). And also clustering methods can be used to generate prototype- or example-based explanations of typical features in the data. Examples here are k-means clustering (Hartigan and Wong 1979) and graph-based spectral clustering (von Luxburg 2007).

Output type

The output type, also considered the actual explanator (Guidotti et al. 2018), describes the type of information presented to the explainee. Note that this ("what" is shown) is mostly independent of the presentation form ("how" it is shown). Typical types are:

By example instance, e.g., closest other samples, word cloud;

Examples The discussed ProtoPNet (Chen et al. 2019a) is based on selecting and comparing relevant example snippets from the input image data.

Contrastive / counterfactual / near miss examples, including adversarial examples;
The goal here is to explain for an input why the respective output was as obtained instead of a desired output. This is done by presenting how the input features have to change in order to obtain the alternative output. Counterfactual examples are sometimes seen as a special case of more general contrastive examples (Stepin et al. 2021). Desirables associated specifcally with counterfactual examples are that they are valid inputs close to the original examples and with few features changed (*sparsity*) that are actionable for the explainee and that they adhere to known causal relations (Guidotti et al. 2021; Verma et al. 2020; Keane et al. 2021).

Examples **CEM** (Dhurandhar et al. 2018). The perturbation-based feature importance heatmapping approach of RISE is extended in CEM (Contrastive, Black-box Explanations Model) by Dhurandhar et al. (2018). They do not only fnd positively contributing features but also the features that must minimally be absent to not change the output.

Prototype, e.g., generated, concept vector;

Examples A typical prototype generator is used in the discussed Feature Visualization method (Olah et al. 2017): images are generated, *e.g.*, via gradient descent, that represent the prototypical pattern for activating a flter . While this considers prototypical inputs, concept embeddings as collected in TCAV (Kim et al. 2018a) and Net2Vec (Fong and Vedaldi 2018) describe prototypical activation patterns for a given semantic concept. The concept mining approach ACE (Ghorbani et al. 2019) combines prototypes with examples: They search a concept embedding as a prototype for an automatically collected set of example patches, that, in turn, can be used to explain the prototype.

Feature importance that will highlight features with high attribution or infuence on the output;

Examples A lot of feature importance methods producing heatmaps have been discussed before, such as RISE (Petsiuk et al. 2018), D-RISE (Petsiuk et al. 2021), CEM (Dhurandhar et al. 2018), LIME (Ribeiro et al. 2016), SHAP (Lundberg and Lee 2017), LRP (Bach et al. 2015), PatternAttribution (Kindermans et al. 2018), Sensitivity Analysis (Baehrens et al. 2010), Deconvnet and successors (Zeiler and Fergus 2014; Simonyan et al. 2014; Springenberg et al. 2015). (Fong and Vedaldi 2017). One further example is the work by Fong and Vedaldi (2017), who follow a perturbation-based approach. Similar to RISE, their idea is to fnd a minimal occlusion mask that, if used to perturb the image (*e.g.*, blur, noise, or blacken), maximally changes the outcome. To fnd the mask, backpropagation is used, making it a model-specifc method.
CAM (Zhou et al. 2016), **Grad-CAM** (Selvaraju et al. 2017). Some older but popular and simpler example methods are Grad-CAM by Selvaraju et al. (2017) and its predecessor CAM (Class Activation Mapping) by Zhou et al. (2016). While Deconvnet and its successors can only consider the feature impor-

tance with respect to intermediate outputs, (Grad-)CAM produces class-specifc heatmaps, which are the weighted sum of the flter activation maps for one (usually the last) convolutional layer. For CAM, it is assumed the convolutional backend is fnalized by a global average pooling layer that densely connects to the fnal classifcation output. Here, the weights in the sum are the weights connecting the neurons of the global average pooling layer to the class outputs. For Grad-CAM, the weights in the sum are the averaged derivation of the class output by each activation map pixel.

Concept-wise Grad-CAM (Zhou et al. 2018). This is also used in the more recent work of Zhou et al. (2018), who do not apply Grad-CAM directly to the output but to each of a minimal set of projections from a convolutional intermediate output of a DNN that predict semantic concepts.

SIDU (Muddamsetty et al. 2021). Similar to Grad-CAM, SIDU (Similarity Distance and Uniqueness) by Muddamsetty et al. (2021) also adds up the flter -wise weighted activations of the last convolutional layer. The weights encompass a combination of a similarity score and a uniqueness score for the prediction output under each flter activation mask. The scores aim for high similarity of a masked prediction with the original one and low similarity to the other masked prediction, leading to masks capturing more interesting object regions.

Rule-based, e.g., decision tree; or if-then, binary, m-of-n, or hyperplane rules (cf. (Hailesilassie 2016));

Examples The mentioned exemplary rule-extraction methods DeepRED (Zilke et al. 2016) and VIA (Thrun 1995), as well as decision tree extractors TREPAN (Craven and Shavlik 1995), Concept Tree (Renard et al. 2019), and C4.5 (Quinlan 1993) all provide global, rule-based output. For further rule extraction examples, we refer the reader to the comprehensive surveys Hailesilassie (2016), Wang et al. (2018a), Augasta and Kathirvalavakumar (2012) on the topic and the survey by Wang et al. (2018b) for recurrent DNNs.

LIME-Aleph (Rabold et al. 2018). An example of a local rule-extractor is the recent LIME-Aleph approach by Rabold et al. (2018), which generates a local explanation in the form of frst-order logic rules. This is learned using inductive logic programming (ILP) (Cropper et al. 2020) trained on the symbolic knowledge about a set of semantically similar examples. Due to the use of ILP, the approach is limited to tabular input data and classifcation outputs, but just like LIME, it is model-agnostic.

NBDT (Wan et al. 2020). A similar approach is followed by NBDT (Neural-Backed Decision Trees). Here, Wan et al. (2020) assume that the concept embeddings of super-categories are represented by the mean of their sub-category vectors (*e.g.*, the mean of "cat" and "dog" should be "animal with four legs"). This is used to infer from bottom-to-top a decision tree where the nodes are super-categories, and the leaves are the classifcation classes. At each node, it is decided which of the sub-nodes best applies to the image. As embedding for a leaf concept (an output class), they suggest taking the weights connecting the penultimate layer to a class output, and as similarity measure for the categories, they use dot-product (cf. Net2Vec and TCAV).

Dimension reduction, i.e., sample points are projected to a sub-space;

Examples Typical dimensionality reduction methods mentioned previously are PCA (Jolliffe 2002) and t-SNE (van der Maaten and Hinton 2008).

Dependence plots which plot the effect of an input feature on the final output of a model (cf. (Adadi and Berrada 2018; Carvalho et al. 2019));

Examples
PDP (Friedman 2001). PDP (Partial Dependency Plots, cf. (Molnar 2020, sec. 5.1)) by Friedman (2001) calculate for one input feature, and for each value of this feature, the expected model outcome averaged over the dataset. This results in a plot (for each output) that indicates the global infuence of the respective feature on the model.
ICE (Goldstein et al. 2015). The local equivalent by Goldstein et al. (2015), ICE (Individual Conditional Expectation, cf. (Molnar 2020, sec. 5.2)) plots, obtain the PDP for generated data samples locally around a given sample.

Graphs as of *e.g.* Mueller et al. (2019, 2021); Linardatos et al. (2021);

Examples The previously discussed Explanatory Graph (Zhang et al. 2018) method provides, amongst others, a graph-based explanation output.

Combinations of mentioned model types.

Presentation

The presentation of information can be characterized by two categories of properties: the used *presentation form* and the *level of abstraction* used to present available information. The presentation form simply summarizes the human sensory input channels utilized by the explanation, which can be: visual (the most common one including diagrams, graphs, and heatmaps), textual in either natural language or formal form, auditive, and combinations thereof. In the following, the aspects infuencing the level of abstraction are elaborated. These can be split up into (1) aspects of the smallest building blocks of the explanation, the *information units*, and (2) the *accessibility* or level of *complexity* of their combinations (the information units). Lastly, further fltering may be applied before fnally presenting the explanation, including privacy flters.

Information units The basic units of the explanation, cognitive chunks (Doshi-Velez and Kim 2017), or information units, may differ in the level of processing applied to them. The simplest form are unprocessed *raw features*, as used in explanations by example. *Derived features* capture some indirect information contained in the raw inputs, like superpixels or attention heatmaps. These need not necessarily have semantic meaning to the explainee, in contrast to explicitly *semantic features*, *e.g.*, concept activation vector attributions. The last type of information unit are *abstract semantic features* not directly grounded in any input, *e.g.*, generated prototypes. *Feature interactions* may occur as information units or be left unconsidered for the explanation.

Examples Some further notable examples of heatmapping methods for feature attribution are SmoothGrad by Smilkov et al. (2017) and Integrated Gradients

by Sundararajan et al. (2017). One drawback of the methods described so far is that they linearly approximate the loss surface in a point-wise manner. Hence, they struggle with "rough" loss surfaces that exhibit signifcant variation in the point-wise values, gradients, and thus feature importance (Samek et al. 2020).

SmoothGrad (Smilkov et al. 2017). SmoothGrad aims to mitigate this by averaging the gradient from random samples within a ball around the sample to investigate.

Integrated Gradients (Sundararajan et al. 2017) Integrated gradients do the averaging (to be precise: integration) along a path between two points in the input space.

Integrated Hessians (Janizek et al. 2020). A technically similar approach but with a different goal is Integrated Hessians (Janizek et al. 2020). They intend not to grasp and visualize the sensitivity of the model for one feature (as a derived feature), but their information units are interactions of features, *i.e.*, how much the change of one feature changes the infuence of the other on the output. This is done by having a look at the Hessian matrix, which is obtained by two subsequent Integrated Gradients calculations.

Accessibility The accessibility, level of detail, or level of complexity describes how much intellectual effort the explainee has to bring up in order to understand the simulatable parts of the explanation. Thus, the perception of complexity heavily depends on the end-user, which is mirrored in the human-grounded complexity / interpretability metric discussed later in Sect. 5.3. In general, one can differentiate between representations that are considered *simpler* and such that are more *expressive but complex*. Because accessibility is a precondition to simulating the parts, it is not the same as the transparency level. For example, very large, transparent decision trees or very high-dimensional (general) linear models may be perceived as globally complex by the end-user. However, when looking at the simulatable parts of the explanator, like small groups of features or nodes, they are easy to grasp.

Examples Accessibility can indirectly be assessed by the complexity and expressivity of the explanation (see Sect. 5.3). To give some examples: Simple presentations are, *e.g.*, linear models, general additive models, decision trees and Boolean decision rules, Bayesian models, or clusters of examples (cf. Sect. 5.1); generally, more complex are, *e.g.*, frst-order or fuzzy logical decision rules.

Privacy awareness Sensible information like names may be contained in parts of the explanation, even though they are not necessary for understanding the actual decision. In such cases, an important point is privacy awareness (Calegari et al. 2020): Is sensible information removed if unnecessary or properly anonymized if needed?

5.2.3 Interactivity

The interaction of the user with the explanator may either be static, so the explainee is once presented with an explanation, or interactive, meaning an iterative process accepting user feedback as explanation input. Interactivity is characterized by the *interaction task* and the *explanation process*.

Interaction task The user can either inspect explanations or *correct* them. Inspecting takes place through *exploration* of different parts of one explanation or through consideration of various alternatives and complementing explanations, such as implemented in the *iNNvestigate* toolbox (Alber et al. 2019). Besides, the user can be empowered within the human-AI partnership to provide corrective feedback to the system via an explanation interface, in order to adapt the explanator and thus the explanandum.

Examples State-of-the-art systems

- enable the user to perform *corrections on labels* and to act upon wrong explanations through interactive machine learning (intML), such as implemented in the approach **CAIPI** (Teso and Kersting 2019),
- they allow for *re-weighting of features* for explanatory debugging, like the system **EluciDebug** (Kulesza et al. 2010),
- *adaption of features* as provided by **Crayons** (Fails and Olsen Jr 2003), and
- correcting generated verbal explanations through user-defned constraints, such as implemented in the medical-decision support system **LearnWith-ME** (Schmid and Finzel 2020).

Explanation process As mentioned above, explanation usually takes place in an iterative fashion. Sequential analysis allows the user to query further information in an iterative manner and to understand the model and its decisions over time, in accordance with the users' capabilities and the given context (El-Assady et al. 2019; Finzel et al. 2021b).

Examples
Multi-modal explanations (Hendricks et al. 2018). The explanation process includes combining different methods to create multi-modal explanations and involving the user through dialogue. It can be realized in a phrase-critic model as presented by Hendricks et al. (2018), or with the help of an explanatory dialogue such as proposed by Finzel et al. (2021b).

5.2.4 Mathematical constraints

Mathematical constraints encode some formal properties of the explanator that were found to be helpful for explanation receival. Constraints mentioned in the literature are:

Linearity Considering a concrete proxy model as explanator output, linearity is often a desirable form of simplicity (Kim et al. 2018a; Carvalho et al. 2019; Molnar 2020).
Monotonicity Similar to linearity, one here considers a concrete proxy model as the output of the explanator. The dependency of that model's output on one input feature may be monotonous. Monotonicity is desirable for the sake of simplicity.
Satisfiability This is the case if the explanator outputs readily allow the application of formal methods like solvers.

Number of iterations While some XAI methods require a one-shot inference of the explanandum model (*e.g.*, gradient-based methods), others require several iterations of queries to the explanandum. Since these might be costly or even restricted in some use cases, a limited number of iterations needed by the explanator may be desirable in some cases. Such restrictions may arise from non-gameability (Langer et al. 2021) constraints on the explanandum model, *i.e.*, the number of queries is restricted in order to guard against systematic optimization of outputs by users (*e.g.*, searching for adversaries).

Size constraints Many explanator types, respectively surrogate model types, allow for architectural constraints, like size or disentanglement, that correlate with reduced complexity. See also the respective complexity/interpretability metrics in Sect. 5.3.

> *Examples* For linear models, sparsity can reduce complexity (Gleicher 2016), and for decision trees, depth and width can be constrained. One common way to achieve sparsity is to add regularization terms to the training of linear explanators and interpretable models.

5.3 Metrics

By now, there is a considerable amount of metrics being suggested to assess the quality of XAI methods with respect to different goals. This section details the types of metrics considered in the literature. Following the original suggestion by Doshi-Velez and Kim (2017), we categorize metrics by their level of human involvement required to measure them:

- *functionally-grounded* metrics,
- *human-grounded* metrics, and
- *application-grounded* metrics.

An overview is given in Fig. 6. For a selection of approaches to measure the below-described metrics, we refer the reader to Li et al. (2020). They provide a good starting point with an in-depth analysis of metrics measurement for visual feature attribution methods.

5.3.1 Functionally-grounded metrics

Metrics are considered functionally-grounded if they do not require any human feedback but instead measure the formal properties of the explanator. This applies to the following metrics:

Faithfulness (Li et al. 2020), fdelity (Carvalho et al. 2019), soundness (Yao 2005), or causality (Calegari et al. 2020), measures how accurately the behavior of the explanator conforms with that of the actual object of explanation. If a full surrogate model is used, this is the accuracy of the surrogate model outputs with respect to the explanandum outputs. And the fdelity of inherently interpretable models also serving as explanators is naturally 100%. Note that fdelity is more often used if the explanator consists of a complete surrogate model (*e.g.*, a linear proxy like LIME (Ribeiro et al. 2016)) and faithfulness in more general contexts (Burkart

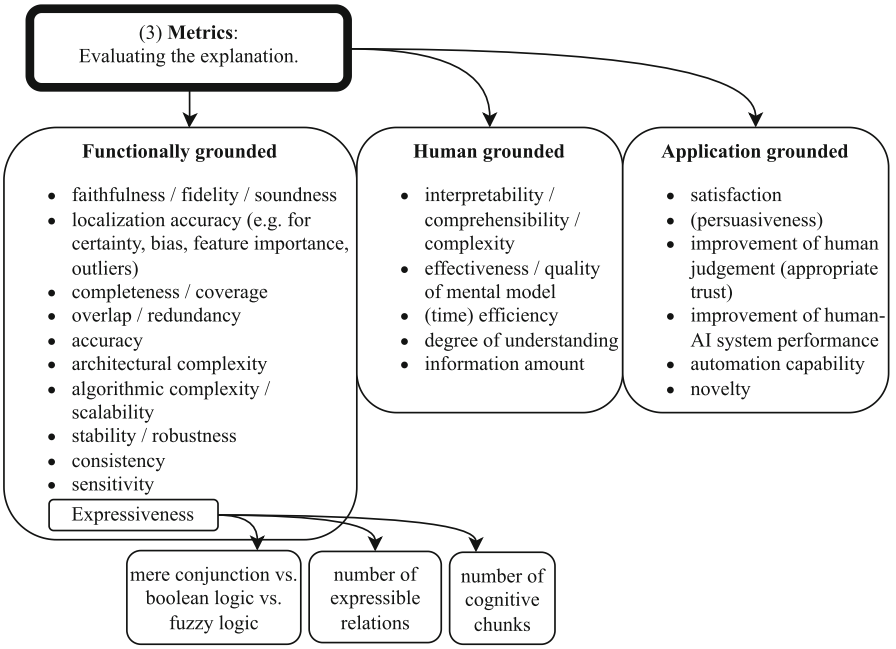

Fig. 6 Overview of the XAI metrics detailed in Sect. 5.3. Find a visualization of the complete taxonomy in Fig. 7

and Huber 2021). Other works use the terms interchangeably like Du et al. (2019); Carvalho et al. (2019), which we adopt here. More simplifcation usually comes along with less faithfulness since corner cases are not captured anymore, also called the *fidelity-interpretability trade-off*.

Localization accuracy with respect to some ground truth (cf. Li et al. 2020 for visual feature importance maps) means how well an explanation correctly localizes certain points of interest. These points of interest must be given by a ground truth that is preferably provided by mathematical properties, such as certainty, bias, feature importance, and outliers (cf. Carvalho et al. 2019). In which way such points are highlighted for the explainee depends on the explanation type. For example, they could be highlighted in a feature importance heatmap (Li et al. 2020) or expressed by the aggregated relevance within regions of interest (Rieger et al. 2020). Note that localization capability is closely related to faithfulness but refers to specifc properties of interest.

Completeness, or coverage, measures how large the validity range of an explanation is, so in which subset of the input space high fdelity can be expected. It can be seen as a generalization of fdelity to the distribution of fdelity . Note that coverage can also be calculated for parts of an explanation, such as single rules of a rule set, as considered by Burkart and Huber (2021).

Overlap is considered by Burkart and Huber (2021) for rule-based explanations. It measures the number of data samples that satisfy more than one rule of the rule set, *i.e.*, measures the size of areas where the validity ranges of different rules overlap.

This can be seen as a measure of redundancy in the rule set and may sometimes correlate with perceived complexity (Burkart and Huber 2021).

Accuracy of the surrogate model ignores the prediction quality of the original model and only measures the prediction quality of the surrogate model for the original task (using the standard accuracy metric). This only applies to post-hoc explanations.

Architectural complexity can be measured using metrics specifc to the explanator type. The goal is to approximate the subjective, human-grounded complexity that a human perceives using purely architectural properties like size measures. Such architectural metrics can be, e.g., the number of used input features for feature importance, the number of changed features for counterfactual examples (also called *sparsity* (Verma et al. 2020)), the sparsity of linear models (Gleicher 2016), the width or depth of decision trees (Burkart and Huber 2021), or, in case of rules, the number of defned rules and the number of unique used predicates (Burkart and Huber 2021).

Algorithmic complexity and scalability measure the information-theoretic complexity of the algorithm used to derive the explanator. This includes the time to convergence (to an acceptable solution) and is especially interesting for complex approximation schemes like rule extraction.

Stability or robustness (Calegari et al. 2020) measures the change of explanator (output) given a change in the input samples. This is an analogon to (adversarial) robustness of deep neural networks and a stable algorithm is usually also better comprehensible and desirable.

Consistency measures the change of the explanator (output) given a change in the model to explain. The idea behind consistency is that functionally equivalent models should produce the same explanation. This assumption is important for model-agnostic approaches, while for model-specifc ones, a dependency on the model architecture may even be desirable. (*e.g.*, for architecture visualization).

Sensitivity measures whether local explanations change if the model output changes strongly. A big change in the model output usually comes along with a change in the discrimination strategy of the model between the differing samples (Li et al. 2020). Such changes should be refected in the explanations. Note that this may be in confict with stability goals for regions in which the explanandum model behaves chaotically.

Expressiveness or the level of detail refers to the level of detail of the formal language used by the explanator. It is interested in approximating the expected information density perceived by the user. It is closely related to the level of abstraction of the presentation. Several functionally-grounded proxies were suggested to obtain comparable measures for expressivity:

- the depth or amount of *added information*, also measured as the mean number of used information units per explanation;
- *number of relations* that can be expressed; and
- the *expressiveness category* of used rules, namely mere conjunction, boolean logic, frst-order logic, or fuzzy rules (cf. (Yao 2005)).

5.3.2 Human-grounded metrics

Other than functionally-grounded metrics, human-grounded metrics require to involve a human directly on proxy tasks for their measurement. Human involvement can be measured through observation of a person's reactions but also through direct human feedback. Often, proxy tasks are considered instead of the fnal application to avoid a need for expensive experts or application runtime (think of medical domains). The goal of an explanation always is that the receiver of the explanation can build a *mental model* of (aspects of) the object of explanation (Kulesza et al. 2013). Human-grounded metrics aim to measure some fundamental psychological properties of the XAI methods, namely quality of the *mental model*. The following are counted as such in literature:

Interpretability or comprehensibility, or complexity measures how accurately the mental model approximates the explanator model. This measure mostly relies on subjective user feedback on whether they "could make sense" of the presented information. It depends on background knowledge, biases, and cognition of the subject and can reveal the use of vocabulary inappropriate to the user (Gilpin et al. 2018).

Effectiveness measures how accurately the mental model approximates the object of explanation. In other words, one is interested in how well a human can simulate the (aspects of interest of the) object after being presented with the explanations. Proxies for effectiveness can be fdelity and accessibility (Molnar 2020, Sec. 2.4). This may serve as a proxy for interpretability.

(Time) efficiency measures how time effcient an explanation is, *i.e.*, how long it takes a user to build up a viable mental model. This is especially of interest in applications with a limited time frame for user reaction, like product recommendation systems (Nunes and Jannach 2017) or automated driving applications (Kim et al. 2018b).

Degree of understanding measures in interactive contexts the current status of understanding. It helps to estimate the remaining time or measures needed to reach the desired extent of the explainee's mental model.

Information amount measures the total subjective amount of information conveyed by one explanation. Even though this may be measured on an information-theoretic basis, it usually is subjective and thus requires human feedback. Functionally-grounded related metrics are the (architectural) complexity of the object of explanation, together with fdelity and coverage. For example, more complex models have a tendency to contain more information, and thus require more complex explanations if they are to be approximated widely and accurately.

5.3.3 Application-grounded metrics

Other than human-grounded metrics, application-grounded ones work on human feedback for the fnal application. The following metrics are considered application-grounded:

Satisfaction measures the direct content of the explainee with the system. It implicitly measures the benefit of explanations for the explanation system user.

Persuasiveness assesses the capability of the explanations to nudge an explainee into a certain direction. This is foremostly considered in recommendation systems (Nunes and Jannach 2017) but has high importance when it comes to analysis tasks, where false positives and false negatives of the human-AI system are undesirable. In this context, a high persuasiveness may indicate a miscalibration of indicativeness.

Improvement of human judgment (Mueller et al. 2021) measures whether the explanation system user develops an appropriate level of trust in the decisions of the explained model. Correct decisions should be trusted more than wrong decisions, *e.g.* because explanations of wrong decisions are illogical.

Improvement of human-AI system performance considers the end-to-end task to be achieved by all of the following: explanandum, explainee, and explanator. This can, *e.g.*, be the diagnosis quality of doctors assisted by a recommendation system (Mueller et al. 2021; Schmid and Finzel 2020).

Automation capability gives an estimate of how much of the manual work conducted by the human in the human-AI system can be automatized. Especially for local explanation techniques, automation may be an important factor for feasibility if the number of samples a human needs to scan can be drastically reduced (Finzel et al. 2021a).

Novelty estimates the subjective degree of novelty of information provided to the explainee (Langer et al. 2021). This is closely related to effcienc y and satisfaction: Especially in exploratory use cases, high novelty can drastically increase effcienc y (no repetitive work for the explainee) and keep satisfaction high (decrease the possibility of boredom for the explainee) (Table 6).

6 Discussion and conclusion

The abundance of existing literature on structuring the feld and methods of XAI has by now reached an overwhelming level for beginners and practitioners. To help in fnding good starting points for a deeper dive, we here presented a rich and structured survey of surveys on XAI topics. This showed an increasing breadth of application felds and method types investigated in order to provide explainable yet accurate learned models. Especially, the exponentially increasing number of both XAI methods and method surveys suggests an increasing interest in XAI and unrestrained growth of the research feld in the upcoming years. As some persisting trends, we found the application domains of medicine and recommendation systems, as well as (visual) explanations for visual models. Some upcoming or re-awakening trends seem to be the feld of natural language processing, and rule-based explanation methods.

With the ever-growing amount of methods for achieving explainability and interpretability, XAI method surveys developed many approaches for method categorization. Early foundational concepts have been extended and sub-structured to increasingly detailed collections of aspects for differentiating XAI methods. This paper systematically united aspects scattered over existing literature into an overarching taxonomy structure. Starting from the defnition of the problem of XAI, we found the following three main parts of an explanation procedure suitable for a top-level taxonomy structure: the task, the explainer, and evaluation metrics. This paper in detail

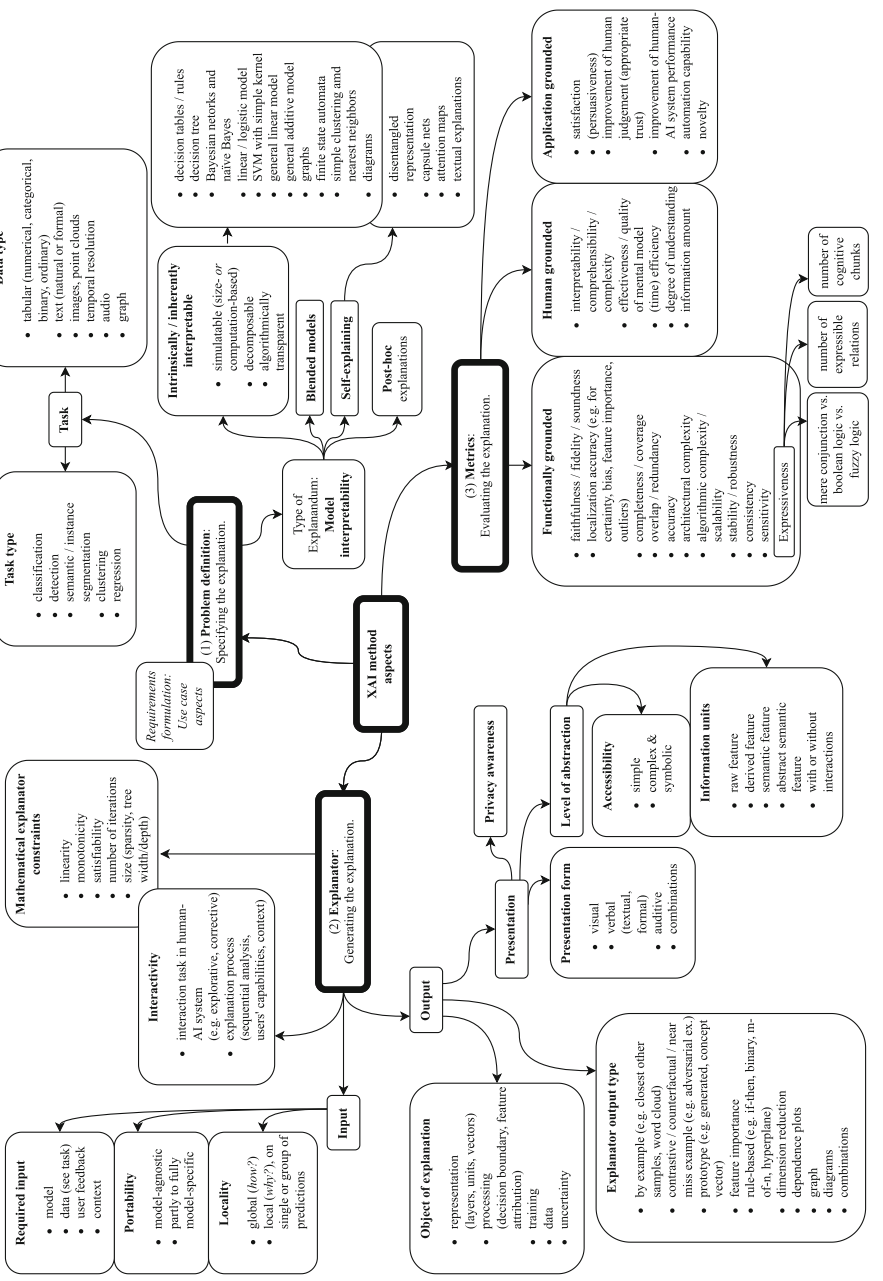

Fig. 7 Overview of the complete taxonomy that is detailed in Sect. 5

Table 6 Review of an exemplary selection of XAI techniques according to the defned taxonomy aspects (without inherently transparent models from Sect. 5.1.2)

Name	Cite	Task	Model-agnostic?	Transp.	Global?	Obj. Expl.	Form	Type
Self-explaining and blended models								
–	Hendricks et al. (2016)	cls		s		p	sym/vis	rules/f
–	Kim et al. (2018b)	any		s		p	sym/vis	rules/f
ProtoPNet	Chen et al. (2019a)	cls,img		s		p/r	vis	proto/f
Capsule Nets	Sabour et al. (2017)	cls		s		r	sym	f
Semantic Bottlenecks, ReNN, Concept Whitening	Losch et al. (2019), Wang (2018), Chen et al. (2020)	any		s		r	sym	f
Logic Tensor Nets	Donadello et al. (2017)	any		b	✓	p/r	sym	rule
FoldingNet	Yang et al. (2017)	any,pcl		b		p	vis	f /red
Neuralized clustering	Kauffmann et al. (2019)	any		b		p	vis	f
Black-box heatmapping								
LIME, SHAP	Ribeiro et al. (2016), Lundberg and Lee (2017)	cls	✓	p		p	vis	f /con
RISE	Petsiuk et al. (2018)	cls,img	✓	p		p	vis	f
D-RISE	Petsiuk et al. (2021)	det,img	✓	p		p	vis	f
CEM	Dhurandhar et al. (2018)	cls,img	✓	p		p	vis	f /con
White-box heatmapping								
Sensitivity analysis	Baehrens et al. (2010)	cls		p		p	vis	f
Deconvnet, (Guided) Backprop.	Zeiler and Fergus (2014), Simonyan et al. (2014), Springenberg et al. (2015)	img		p		p	vis	f
CAM, Grad-CAM	Zhou et al. (2016), Selvaraju et al. (2017)	cls,img		p		p	vis	f
SIDU	Muddamsetty et al. (2021)	cls,img		p		p	vis	f
Concept-wise Grad-CAM	Zhou et al. (2018)	cls,img		p		p/r	vis	f
SIDU	Muddamsetty et al. (2021)	cls,img		p		p	vis	f
LRP	Bach et al. (2015)	cls		p		p	vis	f

Table 6 continued

Name	Cite	Task	Model-agnostic?	Transp.	Global?	Obj. Expl.	Form	Type
Pattern attribution	Kindermans et al. (2018)	cls		p		p	vis	f
–	Fong and Vedaldi (2017)	cls		p		p	vis	f
SmoothGrad, Integrated Gradients	Smilkov et al. (2017), Sundararajan et al. (2017)	cls		p		p	vis	f
Integrated Hessians	Janizek et al. (2020)	cls		p		p	vis	f
Global representation analysis								
Feature Visualization	Olah et al. (2017)	img		p	✓	r	vis	proto
NetDissect	Bau et al. (2017)	img		p	✓	r	vis	proto/f
Net2Vec	Fong and Vedaldi (2018)	img		p	(✓)	r	vis	f
TCAV	Kim et al. (2018a)	any		p	✓	r	vis	f
ACE	Ghorbani et al. (2019)	any		p	✓	r	vis	f
–	Yeh et al. (2020)	any		p	✓	r	vis	proto
IIN	Esser et al. (2020)	any		p	(✓)	r	vis/sym	f
Explanatory Graph	Zhang et al. (2018)	img		p	(✓)	p/r	vis	graph
Dependency plots								
PDP	Friedman (2001)	any	✓	p		p	vis	plt
ICE	Goldstein et al. (2015)	any	✓	p	✓	p	vis	plt
Rule extraction								
TREPAN, C4.5, Concept Tree	Craven and Shavlik (1995), Quinlan (1993), Renard et al. (2019)	cls	✓	p	✓	p	sym	tree
VIA	Thrun (1995)	cls	✓	p	✓	p	sym	rules
DeepRED	Zilke et al. (2016)	cls		p	✓	p	sym	rules
LIME-Aleph	Rabold et al. (2018)	cls	✓	p		p	sym	rules
CA-ILP	Rabold et al. (2020)	cls		p	✓	p	sym	rules
NBDT	Wan et al. (2020)	cls		p	✓	p	sym	tree

Table 6 continued

Name	Cite	Task	Model-agnostic?	Transp.	Global?	Obj. Expl.	Form	Type
Interactivity								
CAIPI	Teso and Kersting (2019)	cls,img	✓	p		r	vis	f /con
EluciDebug	Kulesza et al. (2010)	cls	✓	p		r	vis	f,plt
Crayons	Fails and Olsen Jr (2003)	cls,img	✓	t		p	vis	plt
LearnWithME	Schmid and Finzel (2020)	cls	✓	t	✓	p, r	sym	rules
Multi-modal phrase-critic model	Hendricks et al. (2018)	cls,img		p	✓	p	vis,sym	plt,rules
Inspection of the training								
–	Shwartz-Ziv and Tishby (2017)	any		p	✓	t	vis	dist
Infuence functions	Koh and Liang (2017)	cls		p	✓	t	vis	f /dist
Data analysis methods								
t-SNE, PCA	van der Maaten and Hinton (2008), Jolliffe (2002)	any	✓	p	✓	d	vis	red
k-means, spectral clustering	Hartigan and Wong (1979), von Luxburg (2007)	any	✓	p	✓	d	vis	proto

Abbreviations by column: *image data*=img, *point cloud data*=pcl; *Trans.*=transparency, *post-hoc*=p, *transparent*=t, *self-explaining*=s, *blended*=b; *processing*=p, *representation*=r, *development during training*=t *data*=d; *visual*=vis, *symbolic*=sym, *plot*=plt; *feature importance*=f, *contrastive*=con, *prototypical*=proto, *decision tree*=tree, *distribution*=dist

defnes and sub-structures each of these parts. The applicability of those taxonomy aspects for differentiating methods is evidenced practically: Numerous example methods from the most relevant as well as the most recent research literature are discussed and categorized according to the taxonomy aspects. An overview of the examples is given in the end, concentrating on the seven classifcation criteria that are most significant in the literature. These concretely are the *task*, the form of *interpretability* (*e.g.*, inherently interpretable), whether the method is *model-agnostic or model-specific*, whether it generates *global or local* explanations, what the *object of explanation* is, in what form explanations are *presented*, and the type of explanation.

As highlighted in the course of the paper, the creation of an explanation system should be tailored tightly to the use-case: This holds for all of development, application, and evaluation of XAI the system. Concretely, the different stakeholders and their contexts should be taken into account (Langer et al. 2021; Gleicher 2016). Our proposed taxonomy may serve as a profound basis to analyze stakeholder needs and formulate concrete, use-case-specifc requirements upon the explanation system. The provided survey of surveys then provides the entry point when looking for XAI methods fulflling the derived requirements.

Altogether, our proposed unifed taxonomy and our survey allow (a) beginners to gain an easy and targeted overview and entry point to the feld of XAI; (b) practitioners to formulate suffcient requirements for their explanation use-case and to fnd accordingly suitable XAI methods; and lastly (c) researchers to properly position their research efforts on XAI methods and identify potential gaps. We hope this work fosters research and fruitful application of XAI.

Finally, it must be mentioned that our survey is and will not be or stay complete: Despite our aim for a broad representation of the XAI feld, the used structured literature search is naturally biased by the current interests in specifc sub-felds of AI and the search terms. On the other hand, we hope this ensures relevance to a large part of the research community. Furthermore, we concentrated on a broadly applicable taxonomy for XAI methods, whilst sub-felds of XAI may prefer or come up with more detailed differentiation aspects or a different taxonomy structure.

We are looking forward to seeing future updates of the state of research captured in this work and practical guides for choosing the right XAI method according to a use-case defnition.

Acknowledgements We would like to thank Christian Hellert as well as the anonymous reviewers for their detailed and valuable feedback. The research leading to these results is partly funded by the BMBF ML-3 project Transparent Medical Expert Companion (TraMeExCo), FKZ 01IS18056 B, 2018–2021, and by the German Federal Ministry for Economic Affairs and Energy within the project “KI Wissen – Automotive AI powered by Knowledge”. We would like to thank the consortium for the successful cooperation.

Author contributions No contributors were involved other than the declared authors.

Funding Open Access funding enabled and organized by Projekt DEAL. Funding for this research was obtained as declared in the acknowledgments, with the frst author employed at the Continental Automotive GmbH, Germany, and the second at the University of Bamberg, Germany.

Data/material/code availability Not applicable.

Declarations

Competing interests This work was authored in the scope of doctoral research of both authors, both supervised by Prof. Dr. Ute Schmid at the University of Bamberg. The authors declare that they have no competing financial interests.

Ethics approval and consent to participate Not applicable.

References

Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artifcial intelligence (XAI). In: IEEE Access, pp 52,138–52,160. https://doi.org/10.1109/ACCESS.2018.2870052

Alber M (2019) Software and application patterns for explanation methods. In: Explainable AI: interpreting, explaining and visualizing deep learning. Lecture notes in computer science. Springer, pp 399–433. https://doi.org/10.1007/978-3-030-28954-6_22

Alber M, Lapuschkin S, Seegerer P et al (2019) iNNvestigate neural networks. J Mach Learn Res 20(93):1–8

Allahyari H, Lavesson N (2011) User-oriented assessment of classifcation model understandability. In: 11th Scandinavian conference on artifcial intelligence. IOS Press, http://urn.kb.se/resolve?urn=urn:nbn:se:bth-7559

Altman NS (1992) An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat 46(3):175–185. https://doi.org/10.1080/00031305.1992.10475879

Amershi S, Cakmak M, Knox WB et al (2014) Power to the people: the role of humans in interactive machine learning. AI Mag 35(4):105–120. https://doi.org/10.1609/aimag.v35i4.2513

Ancona M, Ceolini E, Öztireli C, et al (2019) Gradient-based attribution methods. In: Explainable AI: interpreting, explaining and visualizing deep learning. Lecture notes in computer science. Springer, pp 169–191, https://doi.org/10.1007/978-3-030-28954-6_9

Anjomshoae S, Najjar A, Calvaresi D, et al (2019) Explainable agents and robots: results from a systematic literature review. In: 18th international conference autonomous agents and multiagent systems (AAMAS 2019). International Foundation for Autonomous Agents and MultiAgent Systems, pp 1078–1088, http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-158024

Arrieta AB, Rodríguez ND, Ser JD et al (2020) Explainable artifcial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fus. 58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Artelt A, Hammer B (2019) On the computation of counterfactual explanations—a survey. arXiv:1911.07749 [cs, stat]

Arya V, Bellamy RKE, Chen PY, et al (2019) One explanation does not ft all: a toolkit and taxonomy of AI explainability techniques. CoRR arXiv:1909.03012

Augasta MG, Kathirvalavakumar T (2012) Rule extraction from neural networks—a comparative study. In: Proceedings of the 2012 international conference pattern recognition, informatics and medical engineering, pp 404–408. https://doi.org/10.1109/ICPRIME.2012.6208380,

Bach S, Binder A, Montavon G et al (2015) On pixel-wise explanations for non-linear classifer decisions by layer-wise relevance propagation. PLoS ONE 10(7):e0130140. https://doi.org/10.1371/journal.pone.0130140

Baehrens D, Schroeter T, Harmeling S et al (2010) How to explain individual classifcation decisions. J Mach Learn Res 11:1803–1831
Baniecki H, Biecek P (2020) The grammar of interactive explanatory model analysis. arXiv:2005.00497 [Cs Stat]
Bau D, Zhou B, Khosla A, et al (2017) Network dissection: Quantifying interpretability of deep visual representations. In: Proceedings of the 2017 IEEE conference on computer vision and pattern recognition. IEEE Computer Society, pp 3319–3327. https://doi.org/10.1109/CVPR.2017.354
Belle V (2017) Logic meets probability: towards explainable ai systems for uncertain worlds. In: 26th international joint conference on artifcial intelligence, pp 5116–5120
Benchekroun O, Rahimi A, Zhang Q, et al (2020) The need for standardized explainability. arXiv:2010.11273 [Cs]
Biran O, Cotton CV (2017) Explanation and justifcation in machine learning: a survey. In: Proceedings of the IJCAI 2017 workshop explainable artifcial intelligence (XAI)
Bodria F, Giannotti F, Guidotti R, et al (2021) Benchmarking and survey of explanation methods for black box models. arXiv:2102.13076 [cs]
Bruckert S, Finzel B, Schmid U (2020) The next generation of medical decision support: a roadmap toward transparent expert companions. Front Artif Intell 3:75
Burkart N, Huber MF (2021) A survey on the explainability of supervised machine learning. J Artif Intell Res 70:245–317. https://doi.org/10.1613/jair.1.12228
Byrne RMJ (2019) Counterfactuals in explainable artifcial intelligence (XAI): evidence from human reasoning. In: Proceedings of the 2019 international joint conference artifcial intelligence, pp 6276–6282. https://www.ijcai.org/proceedings/2019/876
Calegari R, Ciatto G, Omicini A (2020) On the integration of symbolic and sub-symbolic techniques for XAI: a survey. Intell Artif 14(1):7–32. https://doi.org/10.3233/IA-190036
Carvalho DV, Pereira EM, Cardoso JS (2019) Machine learning interpretability: a survey on methods and metrics. Electronics 8(8):832. https://doi.org/10.3390/electronics8080832
Chang CH, Tan S, Lengerich B, et al (2020) How interpretable and trustworthy are GAMs? CoRR arXiv:2006.06466
Chatzimparmpas A, Martins RM, Jusuf I et al (2020) A survey of surveys on the use of visualization for interpreting machine learning models. Inf Vis 19(3):207–233. https://doi.org/10.1177/1473871620904671
Chen C, Li O, Tao D et al (2019a) This looks like that: deep learning for interpretable image recognition. Adv Neural Inf Process Syst 32:8928–8939
Chen R, Chen H, Huang G, et al (2019b) Explaining neural networks semantically and quantitatively. In: Proceedings of the 2019 IEEE/CVF international conference on computer vision. IEEE, pp 9186–9195. https://doi.org/10.1109/ICCV.2019.00928
Chen Z, Bei Y, Rudin C (2020) Concept whitening for interpretable image recognition. CoRR arXiv:2002.01650
Choudhary P (2018) Interpreting predictive models with skater: unboxing model opacity. O'Reilly Media https://www.oreilly.com/content/interpreting-predictive-models-with-skater-unboxing-model-opacity/
Chromik M, Schüßler M (2020) A taxonomy for human subject evaluation of black-box explanations in XAI. In: Proceedings of the workshop explainable smart systems for algorithmic transparency in emerging technologies, vol 2582. CEUR-WS.org, p 7
Council AUPP (2017) Statement on algorithmic transparency and accountability. Commun ACM
Craven MW, Shavlik JW (1992) Visualizing learning and computation in artifcial neural networks. Int J Artif Intell Tools 1(03):399–425
Craven MW, Shavlik JW (1995) Extracting tree-structured representations of trained networks. In: Advances in Neural Information Processing Systems 8, NIPS, Denver, CO, USA, November 27–30, 1995. MIT Press, pp 24–30, http://papers.nips.cc/paper/1152-extracting-tree-structured-representations-of-trained-networks
Cropper A, Dumancic S, Muggleton SH (2020) Turning 30: new ideas in inductive logic programming. CoRR arXiv:2002.11002
Danilevsky M, Qian K, Aharonov R, et al (2020) A survey of the state of explainable ai for natural language processing. arXiv:2010.00711 [cs]
Das A, Rad P (2020) Opportunities and challenges in explainable artifcial intelligence (XAI): a survey. arXiv:2006.11371

Dey AK (2001) Understanding and using context. Pers Ubiquitous Comput 5(1):4–7. https://doi.org/10.1007/s007790170019

Dhurandhar A, Chen PY, Luss R, et al (2018) Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In: Advances in neural information processing systems 31. Curran Associates, Inc., pp 592–603, https://proceedings.neurips.cc/paper/2018/file/c5ff2543b53f4cc0ad3819a36752467b-Paper.pdf

Donadello I, Serafni L, d'Avila Garcez AS (2017) Logic tensor networks for semantic image interpretation. In: Proceedings of the 26th international joint conference on artifcial intelligence. ijcai.org, pp 1596–1602. https://doi.org/10.24963/ijcai.2017/221

Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv e-prints abs/1702.08608

Došilović FK, Brcić M, Hlupić N (2018) Explainable artifcial intelligence: A survey. In: 2018 41st international convention on information and communication technology, electronics and microelectronics (MIPRO), pp 0210–0215. https://doi.org/10.23919/MIPRO.2018.8400040

Du M, Liu N, Hu X (2019) Techniques for interpretable machine learning. Commun ACM 63(1):68–77. https://doi.org/10.1145/3359786

El-Assady M, Jentner W, Kehlbeck R, et al (2019) Towards xai: structuring the processes of explanations. In: ACM workshop on human-centered machine learning

Esser P, Rombach R, Ommer B (2020) A disentangling invertible interpretation network for explaining latent representations. In: Proceedings 2020 IEEE conference on computer vision and pattern recognition. IEEE, pp 9220–9229. https://doi.org/10.1109/CVPR42600.2020.00924, https://openaccess.thecvf.com/content_CVPR_2020/papers/Esser_A_Disentangling_Invertible_Interpretation_Network_for_Explaining_Latent_Representations_CVPR_2020_paper.pdf

Fails JA, Olsen Jr DR (2003) Interactive machine learning. In: Proceedings of the 8th international conference on Intelligent user interfaces, pp 39–45

Ferreira JJ, Monteiro MS (2020) What are people doing about XAI user experience? A survey on AI explainability research and practice. In: Design, user experience, and usability. Design for contemporary interactive environments. Lecture notes in computer science. Springer, pp 56–73. https://doi.org/10.1007/978-3-030-49760-6_4

Finzel B, Kollmann R, Rieger I, et al (2021a) Deriving temporal prototypes from saliency map clusters for the analysis of deep-learning-based facial action unit classifcation. In: Seidl T, Fromm M, Obermeier S (eds) Proceedings of the LWDA 2021 Workshops: FGWM, KDML, FGWI-BIA, and FGIR, Online, September 1–3, 2021, CEUR Workshop Proceedings, vol 2993. CEUR-WS.org, pp 86–97, http://ceur-ws.org/Vol-2993/paper-09.pdf

Finzel B, Tafer DE, Scheele S et al (2021b) Explanation as a process: user-centric construction of multi-level and multi-modal explanations. In: Edelkamp S, Möller R, Rueckert E (eds) KI 2021: advances in artifcial intelligence. Springer, Cham, pp 80–94

Fong R, Vedaldi A (2018) Net2Vec: Quantifying and explaining how concepts are encoded by flters in deep neural networks. In: Proceedings of the 2018 IEEE conference on computer vision and pattern recognition. IEEE Computer Society, pp 8730–8738. https://doi.org/10.1109/CVPR.2018.00910

Fong RC, Vedaldi A (2017) Interpretable explanations of black boxes by meaningful perturbation. In: Proceedings of the 2017 IEEE international conference on computer vision. IEEE Computer Society, pp 3449–3457. https://doi.org/10.1109/ICCV.2017.371, arXiv:1704.03296

Freitas AA (2014) comprehensible classifcation models: a position paper. ACM SIGKDD Explor Newsl 15(1):1–10. https://doi.org/10.1145/2594473.2594475

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232. https://doi.org/10.1214/aos/1013203451

Ghorbani A, Wexler J, Zou JY et al (2019) Towards automatic concept-based explanations. Adv Neural Inf Process Syst 32:9273–9282

Gilpin LH, Bau D, Yuan BZ, et al (2018) Explaining explanations: an overview of interpretability of machine learning. In: Proceedings if the 5th IEEE international conference on data science and advanced analytics. IEEE, pp 80–89. https://doi.org/10.1109/DSAA.2018.00018

Gleicher M (2016) A framework for considering comprehensibility in modeling. Big Data 4(2):75–88. https://doi.org/10.1089/big.2016.0007

Goebel R, Chander A, Holzinger K, et al (2018) Explainable AI: the new 42? In: Machine learning and knowledge extraction. Lecture notes in computer science. Springer, pp 295–303.https://doi.org/10.1007/978-3-319-99740-7_21

Goldstein A, Kapelner A, Bleich J et al (2015) Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Stat 24(1):44–65. https://doi.org/10.1080/10618600.2014.907095
Goodman B, Flaxman S (2017) European union regulations on algorithmic decision-making and a "right to explanation". AI Mag 38(3):50–57. https://doi.org/10.1609/aimag.v38i3.2741
Guidotti R (2022) Counterfactual explanations and how to fnd them: literature review and benchmarking. Data Min Knowl Disc. https://doi.org/10.1007/s10618-022-00831-6
Guidotti R, Monreale A, Ruggieri S et al (2018) A survey of methods for explaining black box models. ACM Comput Surv 51(5):931–9342. https://doi.org/10.1145/3236009
Guidotti R, Monreale A, Pedreschi D, et al (2021) Principles of explainable artifcial intelligence. In: Explainable AI within the digital transformation and cyber physical systems: XAI methods and applications. Springer, pp 9–31. https://doi.org/10.1007/978-3-030-76409-8_2,
Gunning D, Aha D (2019) Darpa's explainable artifcial intelligence (xai) program. AI Mag 40(2):44–58
Gunning D, Stefk M, Choi J et al (2019) XAI–explainable artifcial intelligence. Sci Robot. https://doi.org/10.1126/scirobotics.aay7120
Hailesilassie T (2016) Rule extraction algorithm for deep neural networks: a review. arXiv:1610.05267
Hartigan JA, Wong MA (1979) Algorithm AS 136: a k-means clustering algorithm. J R Stat Soc Ser C Appl Stat 28(1):100–108. https://doi.org/10.2307/2346830
Hendricks LA, Akata Z, Rohrbach M, et al (2016) Generating visual explanations. In: Computer vision–ECCV 2016. Lecture notes in computer science. Springer, pp 3–19. https://doi.org/10.1007/978-3-319-46493-0_1
Hendricks LA, Hu R, Darrell T, et al (2018) Grounding visual explanations. In: Proceedings of the European conference on computer vision (ECCV), pp 264–279
Henne M, Schwaiger A, Roscher K, et al (2020) Benchmarking uncertainty estimation methods for deep learning with safety-related metrics. In: Proceedings of the workshop artifcial intelligence safety, CEUR workshop proceedings, vol 2560. CEUR-WS.org, pp 83–90, http://ceur-ws.org/Vol-2560/paper35.pdf
Heuillet A, Couthouis F, Díaz-Rodríguez N (2021) Explainability in deep reinforcement learning. Knowl-Based Syst 214(106):685. https://doi.org/10.1016/j.knosys.2020.106685
Huysmans J, Dejaeger K, Mues C et al (2011) An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. Decis Support Syst 51(1):141–154. https://doi.org/10.1016/j.dss.2010.12.003
Islam SR, Eberle W, Ghafoor SK, et al (2021) Explainable artifcial intelligence approaches: a survey. arXiv:2101.09429
ISO/TC 22 Road vehicles (2020) ISO/TR 4804:2020: road vehicles—safety and cybersecurity for automated driving systems—design, verifcation and validation, 1st edn. International Organization for Standardization, https://www.iso.org/standard/80363.html
ISO/TC 22/SC 32 (2018) ISO 26262-6:2018(En): road vehicles—functional safety—Part 6: product development at the software level, ISO 26262:2018(En), vol 6, 2nd edn. International Organization for Standardization, https://www.iso.org/standard/68388.html
Jackson P (1998) Introduction to expert systems, 3rd edn. Addison-Wesley Longman Publishing Co. Inc, New York
Janizek JD, Sturmfels P, Lee SI (2020) Explaining explanations: axiomatic feature interactions for deep networks. CoRR arXiv:2002.04138
Jolliffe IT (2002) Principal component analysis, 2nd edn. Springer Series in Statistics. Springer, https://doi.org/10.1007/b98835
Karimi AH, Barthe G, Schölkopf B, et al (2021) A survey of algorithmic recourse: defnitions, formulations, solutions, and prospects. arXiv:2010.04050 [cs, stat]
Kauffmann J, Esders M, Montavon G, et al (2019) From clustering to cluster explanations via neural networks. arXiv:1906.07633 [cs, stat]
Keane MT, Kenny EM, Delaney E, et al (2021) If only we had better counterfactual explanations: fv e key defcits to rectify in the evaluation of counterfactual xai techniques. In: Twenty-ninth international joint conference on artifcial intelligence, pp 4466–4474. https://doi.org/10.24963/ijcai.2021/609
Kendall A, Gal Y (2017) What uncertainties do we need in Bayesian deep learning for computer vision? Adv Neural Inf Process Syst 30:5580–5590

Kim J, Canny JF (2017) Interpretable learning for self-driving cars by visualizing causal attention. In: Proceedings of the 2017 IEEE international conference on computer vision. IEEE Computer Society, pp 2961–2969. https://doi.org/10.1109/ICCV.2017.320

Kim B, Wattenberg M, Gilmer J, et al (2018a) Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). In: Proceedings of the 35th international conference on machine learning, proceedings of machine learning research, vol 80. PMLR, pp 2668–2677, http://proceedings.mlr.press/v80/kim18d.html

Kim J, Rohrbach A, Darrell T, et al (2018b) Textual explanations for self-driving vehicles. In: Proc. 15th European conference on computer vision, Part II, Lecture notes in computer science, vol 11206. Springer, pp 577–593. https://doi.org/10.1007/978-3-030-01216-8_35, arXiv:1807.11546

Kindermans PJ, Schütt KT, Alber M, et al (2018) Learning how to explain neural networks: PatternNet and PatternAttribution. In: Proceedings of the 6th international conference on learning representations, https://openreview.net/forum?id=Hkn7CBaTW

Klaise J, Looveren AV, Vacanti G et al (2021) Alibi explain: algorithms for explaining machine learning models. J Mach Learn Res 22(181):1–7

Koh PW, Liang P (2017) Understanding Black-box Predictions via Influence Functions. In: Proceedings of the 34th international conference on machine learning. PMLR, pp 1885–1894, http://proceedings.mlr.press/v70/koh17a.html

Kulesza T, Stumpf S, Burnett M, et al (2010) Explanatory debugging: supporting end-user debugging of machine-learned programs. In: 2010 IEEE symposium on visual languages and human-centric computing. IEEE, pp 41–48

Kulesza T, Stumpf S, Burnett M, et al (2013) Too much, too little, or just right? Ways explanations impact end users' mental models. In: 2013 IEEE symposium on visual languages and human centric computing. IEEE, pp 3–10

Kulesza T, Burnett M, Wong WK, et al (2015) Principles of explanatory debugging to personalize interactive machine learning. In: Proceedings of the 20th international conference intelligent user interfaces, pp 126–137

Langer M, Oster D, Speith T, et al (2021) What do we want from explainable artifcial intelligence (XAI)?—A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artif Intell, p 103473. https://doi.org/10.1016/j.artint.2021.103473

Lapuschkin S, Wäldchen S, Binder A et al (2019) Unmasking Clever Hans predictors and assessing what machines really learn. Nat Commun 10(1):1096. https://doi.org/10.1038/s41467-019-08987-4

Li XH, Shi Y, Li H, et al (2020) Quantitative evaluations on saliency methods: an experimental study. arXiv:2012.15616

Linardatos P, Papastefanopoulos V, Kotsiantis S (2021) Explainable AI: a review of machine learning interpretability methods. Entropy 23(1):18. https://doi.org/10.3390/e23010018

Lipton ZC (2018) The mythos of model interpretability. Queue 16(3):31–57. https://doi.org/10.1145/3236386.3241340

Losch M, Fritz M, Schiele B (2019) Interpretability beyond classifcation output: semantic bottleneck networks. In: Proceedings of the 3rd ACM computer science in cars symposium extended abstracts, https://arxiv.org/pdf/1907.10882.pdf

Lundberg SM, Lee SI (2017) A unifed approach to interpreting model predictions. Adv Neural Inf Process Syst 30:4765–4774

Mazzine R, Martens D (2021) A framework and benchmarking study for counterfactual generating methods on tabular data. arXiv:2107.04680 [cs]

McAllister R, Gal Y, Kendall A, et al (2017) Concrete problems for autonomous vehicle safety: advantages of Bayesian deep learning. In: Proceedings of the 26th international joint conference artifcial intelligence, pp 4745–4753, https://doi.org/10.24963/ijcai.2017/661

McCarthy J (1958) Programs with common sense. In: Proceedings of the Teddington conference on the mechanisation of thought processes, pp 77–84

Miller T (2019) Explanation in artifcial intelligence: insights from the social sciences. Artif Intell 267:1–38. https://doi.org/10.1016/j.artint.2018.07.007

Molnar C (2020) Interpretable machine learning. Lulu.com, https://christophm.github.io/interpretable-ml-book/

Muddamsetty SM, Jahromi MNS, Ciontos AE, et al (2021) Introducing and assessing the explainable AI (XAI) method: SIDU. arXiv:2101.10710

Mueller ST, Hoffman RR, Clancey W, et al (2019) Explanation in human-AI systems: a literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. arXiv:1902.01876

Mueller ST, Veinott ES, Hoffman RR, et al (2021) Principles of explanation in human-AI systems. CoRR arXiv:2102.04972

Muggleton SH, Schmid U, Zeller C et al (2018) Ultra-strong machine learning: comprehensibility of programs learned with ilp. Mach Learn 107(7):1119–1140

Murdoch WJ, Singh C, Kumbier K et al (2019) Defnitions, methods, and applications in interpretable machine learning. PNAS 116(44):22071–22080. https://doi.org/10.1073/pnas.1900654116

Nguyen A, Yosinski J, Clune J (2019) Understanding neural networks via feature visualization: a survey. In: Explainable AI: interpreting, explaining and visualizing deep learning. Lecture notes in computer science. Springer, pp 55–76. https://doi.org/10.1007/978-3-030-28954-6_4

Nori H, Jenkins S, Koch P, et al (2019) InterpretML: a unifed framework for machine learning interpretability. CoRR arXiv:1909.09223

Nunes I, Jannach D (2017) A systematic review and taxonomy of explanations in decision support and recommender systems. User Model User Adap Inter 27(3–5):393–444. https://doi.org/10.1007/s11257-017-9195-0

Olah C, Mordvintsev A, Schubert L (2017) Feature visualization. Distill 2(11):e7. https://doi.org/10.23915/distill.00007

Páez A (2019) The pragmatic turn in explainable artifcial intelligence (xai). Mind Mach 29(3):441–459

Petsiuk V, Das A, Saenko K (2018) RISE: randomized input sampling for explanation of black-box models. In: Proceedings of the British machine vision conference. BMVA Press, p 151, http://bmvc2018.org/contents/papers/1064.pdf

Petsiuk V, Jain R, Manjunatha V, et al (2021) Black-box explanation of object detectors via saliency maps. In: Proceedings of the 2021 IEEE/CVF conference on computer vision and pattern recognition, pp 11443–11452, https://openaccess.thecvf.com/content/CVPR2021/html/Petsiuk_Black-Box_Explanation_of_Object_Detectors_via_Saliency_Maps_CVPR_2021_paper.html

Pocevičiūtė M, Eilertsen G, Lundström C (2020) Survey of XAI in digital pathology. Lect Notes Comput Sci 2020:56–88. https://doi.org/10.1007/978-3-030-50402-1_4

Puiutta E, Veith EMSP (2020) Explainable reinforcement learning: a survey. In: Machine learning and knowledge extraction—4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 international cross-domain conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings, lecture notes in computer science, vol 12279. Springer, pp 77–95, https://doi.org/10.1007/978-3-030-57321-8_5

Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann series in machine learning. Morgan Kaufmann, https://kupdf.net/download/j-ross-quinlan-c4-5-programs-for-machine-learning-1993_5b095daee2b6f5024deefc30_pdf

Rabold J, Siebers M, Schmid U (2018) Explaining black-box classifers with ILP—empowering LIME with Aleph to approximate non-linear decisions with relational rules. In: International conference on machine inductive logic programming. Lecture notes in computer science. Springer, pp 105–117. https://doi.org/10.1007/978-3-319-99960-9_7,

Rabold J, Schwalbe G, Schmid U (2020) Expressive explanations of DNNs by combining concept analysis with ILP. In: KI 2020: advances in artifcial intelligence. Lecture notes in computer science. Springer, pp 148–162. https://doi.org/10.1007/978-3-030-58285-2_11

Renard X, Woloszko N, Aigrain J, et al (2019) Concept tree: high-level representation of variables for more interpretable surrogate decision trees. In: Proceedings of the 2019 ICML workshop human in the loop learning, arXiv:1906.01297

Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?": explaining the predictions of any classifer . In: Proceedings of the 22nd ACM SIGKDD International conference on knowledge discovery and data mining. ACM, KDD'16, pp 1135–1144, arXiv:1602.04938

Rieger I, Kollmann R, Finzel B, et al (2020) Verifying deep learning-based decisions for facial expression recognition. In: Proceedings of the ESANN conference 2020

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215. https://doi.org/10.1038/s42256-019-0048-x

Sabour S, Frosst N, Hinton GE (2017) Dynamic routing between capsules. Adv Neural Inf Process Syst 30:3856–3866

Saeed W, Omlin C (2021) Explainable AI (XAI): a systematic meta-survey of current challenges and future opportunities. https://doi.org/10.48550/arXiv.2111.06420

Samek W, Müller KR (2019) Towards explainable artifcial intelligence. In: Explainable AI: interpreting, explaining and visualizing deep learning, Lecture notes in computer science, vol 11700. Springer, p 5–22. https://doi.org/10.1007/978-3-030-28954-6_1

Samek W, Montavon G, Vedaldi A et al (2019) Explainable AI: interpreting, explaining and visualizing deep learning, Lecture notes in computer science, vol 11700. Springer. https://doi.org/10.1007/978-3-030-28954-6

Samek W, Montavon G, Lapuschkin S, et al (2020) Toward interpretable machine learning: transparent deep neural networks and beyond. CoRR arXiv:2003.07631

Schmid U, Finzel B (2020) Mutual explanations for cooperative decision making in medicine. KI-Künstliche Intelligenz pp 227–233

Schmid U, Zeller C, Besold T, et al (2016) How does predicate invention affect human comprehensibility? In: International conference on inductive logic programming. Springer, pp 52–67

Selvaraju RR, Cogswell M, Das A, et al (2017) Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the 2017 IEEE international conference on computer vision. IEEE, pp 618–626, https://doi.org/10.1109/ICCV.2017.74, https://arxiv.org/abs/1610.02391

Shwartz-Ziv R, Tishby N (2017) Opening the black box of deep neural networks via information. CoRR arXiv:1703.00810

Simonyan K, Vedaldi A, Zisserman A (2014) Deep inside convolutional networks: visualising image classifcation models and saliency maps. In: Proceedings of the 2nd international conference on learning representations, workshop track proceedings, arXiv:1312.6034

Singh A, Sengupta S, Lakshminarayanan V (2020) Explainable deep learning models in medical image analysis. J Imaging 6(6):52. https://doi.org/10.3390/jimaging6060052

Smilkov D, Thorat N, Kim B, et al (2017) SmoothGrad: removing noise by adding noise. CoRR arXiv:1706.03825

Sokol K, Hepburn A, Poyiadzi R et al (2020) Fat forensics: a python toolbox for implementing and deploying fairness, accountability and transparency algorithms in predictive systems. J Open Source Softw 5(49):1904. https://doi.org/10.21105/joss.01904

Spinner T, Schlegel U, Schafer H et al (2020) explAIner: a visual analytics framework for interactive and explainable machine learning. IEEE Trans Vis Comput Gr 26:1064–1074. https://doi.org/10.1109/TVCG.2019.2934629

Springenberg JT, Dosovitskiy A, Brox T, et al (2015) Striving for simplicity: the all convolutional net. In: Proceedings of the 3rd international conference on learning representations, ICLR 2015, workshop track proceedings, arXiv:1412.6806

Stepin I, Alonso JM, Catala A et al (2021) A survey of contrastive and counterfactual explanation generation methods for explainable artifcial intelligence. IEEE Access 9:11974–12001. https://doi.org/10.1109/ACCESS.2021.3051315

Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. In: Proceedings of the 34th international conference on machin learning, proceedings of machine learning research, vol 70. PMLR, pp 3319–3328, http://proceedings.mlr.press/v70/sundararajan17a.html

Teso S, Kersting K (2019) Explanatory interactive machine learning. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society, pp 239–245

Thrun S (1995) Extracting rules from artifcial neural networks with distributed representations. Adv Neural Inf Process Syst 7:505–512

Tintarev N, Masthoff J (2007) A survey of explanations in recommender systems. In: IEEE 23rd international conference on data engineering workshop, pp 801–810. https://doi.org/10.1109/ICDEW.2007.4401070

Tjoa E, Guan C (2020) A survey on explainable artifcial intelligence (XAI): toward medical XAI. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2020.3027314

van der Maaten L, Hinton G (2008) Visualizing Data using t-SNE. J Mach Learn Res 9(86):2579–2605

van Lent M, Fisher W, Mancuso M (2004) An explainable artifcial intelligence system for small-unit tactical behavior. In: Proceedings of the 2004 national conference artifcial intelligence. AAAI Press; 1999, pp 900–907

Vassiliades A, Bassiliades N, Patkos T (2021/ed) Argumentation and explainable artifcial intelligence: a survey. Knowl Eng Rev. https://doi.org/10.1017/S0269888921000011

Verma S, Dickerson J, Hines K (2020) Counterfactual explanations for machine learning: a review. arXiv:2010.10596 [cs, stat]

Vilone G, Longo L (2020) Explainable artifcial intelligence: a systematic review. arXiv:2006.00093

von Luxburg U (2007) A tutorial on spectral clustering. Stat Comput 17(4):395–416. https://doi.org/10.1007/s11222-007-9033-z
Wan A, Dunlap L, Ho D, et al (2020) NBDT: neural-backed decision tree. In: Posters 2021 international conference on learning representations, https://openreview.net/forum?id=mCLVeEpplNE
Wang H (2018) ReNN: rule-embedded neural networks. In: Proceedings of the 24th international conference on pattern recognition. IEEE Computer Society, pp 824–829. https://doi.org/10.1109/ICPR.2018.8545379, http://arxiv.org/abs/1801.09856
Wang Q, Zhang K, II AGO, et al (2018a) A comparative study of rule extraction for recurrent neural networks. CoRR arXiv:1801.05420
Wang Q, Zhang K, Ororbia AG II et al (2018b) An empirical evaluation of rule extraction from recurrent neural networks. Neural Comput 30(9):2568–2591. https://doi.org/10.1162/neco_a_01111
Weitz K (2018) Applying explainable artifcial intelligence for deep learning networks to decode facial expressions of pain and emotions. Master's thesis, Otto-Friedrich-University Bamberg, http://www.cogsys.wiai.uni-bamberg.de/theses/weitz/Masterarbeit_Weitz.pdf
Xie N, Ras G, van Gerven M, et al (2020) Explainable deep learning: a feld guide for the uninitiated. CoRR arXiv:2004.14545
Yang Y, Feng C, Shen Y, et al (2017) Foldingnet: interpretable unsupervised learning on 3d point clouds. CoRR arXiv:1712.07262
Yao J (2005) Knowledge extracted from trained neural networks: What's next? In: Data mining, intrusion detection, information assurance, and data networks security 2005, Orlando, Florida, USA, March 28–29, 2005. SPIE Proceedings, vol 5812. SPIE, pp 151–157. https://doi.org/10.1117/12.604463
Yeh CK, Kim B, Arik S et al (2020) On completeness-aware concept-based explanations in deep neural networks. Adv Neural Inf Process Syst 33:20554–20565
Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: Proceedings of the 13th European conference on computer vision—part I, lecture notes in computer science, vol 8689. Springer, pp 818–833, https://doi.org/10.1007/978-3-319-10590-1_53,
Zhang Q, Zhu SC (2018) Visual interpretability for deep learning: a survey. Front IT EE 19(1):27–39. https://doi.org/10.1631/FITEE.1700808
Zhang Y, Chen X (2020) Explainable recommendation: a survey and new perspectives. FNT Inf Retr 14(1):1–101. https://doi.org/10.1561/1500000066
Zhang Q, Cao R, Shi F, et al (2018) Interpreting CNN knowledge via an explanatory graph. In: Proceedings of the 32nd AAAI conference on artifcial intelligence. AAAI Press, pp 4454–4463, https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17354
Zhang Y, Tino P, Leonardis A, Tang K (2021) A survey on neural network interpretability. IEEE Trans Emerg Top Comput Intell 5(5):726–742. https://doi.org/10.1109/TETCI.2021.3100641
Zhou B, Khosla A, Lapedriza À, et al (2016) Learning deep features for discriminative localization. In: Proceedings of the 2016 IEEE conference computer vision and pattern recognition. IEEE Computer Society, pp 2921–2929. https://doi.org/10.1109/CVPR.2016.319, arXiv:1512.04150
Zhou B, Sun Y, Bau D, et al (2018) Interpretable basis decomposition for visual explanation. In: Computer vision—ECCV 2018. Lecture notes in computer science. Springer, pp 122–138. https://doi.org/10.1007/978-3-030-01237-3_8
Zhou J, Gandomi AH, Chen F et al (2021) Evaluating the quality of machine learning explanations: a survey on methods and metrics. Electronics 10(5):593. https://doi.org/10.3390/electronics10050593
Zilke JR, Loza Mencía E, Janssen F (2016) DeepRED—rule extraction from deep neural networks. In: Proceedings of the 19th international conference discovery science, Lecture notes in computer science. Springer, pp 457–473. https://doi.org/10.1007/978-3-319-46307-0_29,