A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods

Timo Speith
timo.speith@uni-saarland.de
Saarland University
Saarbrücken, Saarland, Germany

ABSTRACT

The recent surge in publications related to explainable artificial intelligence (XAI) has led to an almost insurmountable wall if one wants to get started or stay up to date with XAI. For this reason, articles and reviews that present taxonomies of XAI methods seem to be a welcomed way to get an overview of the field. Building on this idea, there is currently a trend of producing such taxonomies, leading to several competing approaches to construct them. In this paper, we will review recent approaches to constructing taxonomies of XAI methods and discuss general challenges concerning them as well as their individual advantages and limitations. Our review is intended to help scholars be aware of challenges current taxonomies face. As we will argue, when charting the field of XAI, it may not be sufficient to rely on one of the approaches we found. To amend this problem, we will propose and discuss three possible solutions: a new taxonomy that incorporates the reviewed ones, a database of XAI methods, and a decision tree to help choose fitting methods.

CCS CONCEPTS

• **General and reference** → **Surveys and overviews**; • **Computing methodologies** → *Artificial intelligence*.

KEYWORDS

explainability, interpretability, explainable artificial intelligence, XAI, transparency, taxonomy, review

ACM Reference Format:
Timo Speith. 2022. A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3531146.3534639

1 INTRODUCTION

As artificial intelligence (AI) advances into more and more sensitive areas of daily life, such as healthcare [34] and criminal justice [25], questions about the ethics, fairness, and safety of AI-based systems become increasingly pressing [58]. However, the opaque nature of state-of-the-art machine learning (ML) techniques, such as deep neural networks (DNNs) and random forests, makes it challenging to comprehend the inner workings of learned models [18, 20].

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3534639

To amend the lack of understanding of AI-based systems, their reasoning processes, and their outputs, the research field of explainable AI (XAI) re-emerged in recent years [21, 40, 44, 57] after its genesis in the 80s [16, 22]. However, understanding (some aspect of) a system is just an intermediary step to other goals, such as those mentioned above [10, 21, 44]. By understanding how a particular system's output (e.g., a denied loan decision) came to be, a person should be enabled to assess whether this output was based on valid criteria or not [29, 57]. If, for instance, an unfavorable decision was based on (a proxy for) the skin color of a person, this person should be able to recognize that they were treated unfairly [19, 41].

As a research discipline, XAI has grown exponentially in recent years, with annual publications easily exceeding several hundreds [10, 75]. Accordingly, not only is it hard for new scholars to get started with research in XAI, even for experienced scholars it is challenging to keep track of new developments and trends. To amend these problems, there is an increasing number of reviews and overview articles on specific facets of XAI. One especially popular practice is building taxonomies of proposed methods in XAI (we will call these methods *explainability methods*). Solely for the year 2021, we are aware of six papers that present some form of taxonomy of explainability methods: [6, 14, 44, 49, 54, 72].

While taxonomies can serve as a useful tool to organize debates, the landscape of explainability methods is simply too broad and complex to be compressed into a single pragmatically useful taxonomy. Accordingly, each taxonomy inevitably has to focus on certain aspects and leave out others if it is to be pragmatically useful. Consequently, there are several, distinct approaches to constructing taxonomies of explainability methods. However, this diversity gives rise to several challenges for researchers constructing taxonomies as well as recipients utilizing them.

On the one hand, *researchers* may have problems coming up with appropriate and representative classification categories within a single taxonomy, especially if these are intended to be distinct. For instance, presented categories are often depicted as mutually exclusive (i.e., one method can only belong to a single category), whereas, in reality, they are not: in principle, one method could often be assigned to several categories simultaneously. Here, we stress that researchers should strive for pragmatically useful taxonomies that avoid misrepresentation of the field. In particular, researchers should become aware of potential overlaps in classification, take them into account, and explicitly acknowledge them.

On the other hand, *recipients* may struggle with a vast amount of taxonomies that lack uniformity between each other. Owing to the problems researchers face, these taxonomies differ with respect to the categories of explainability methods they propose. Furthermore, even when they have roughly comparable classification categories,

the same method is sometimes sorted into different ones across the various taxonomies. By raising awareness of these challenges, we aim to help recipients navigate the debate.

While it is only reasonable that taxonomies with different foci have some discrepancies,[1] such differences may a) confuse people new to the field of XAI and b) prevent them from acquiring an adequate picture of the XAI landscape. To provide a remedy, we will examine several recent approaches to constructing taxonomies of explainability methods and discuss general challenges, as well as individual advantages and limitations. Among other purposes, this paper is intended to help newcomers to the field of XAI find the taxonomy that is most conducive to achieving the specific goal(s) they have. Furthermore, experts in XAI can also take away valuable insights by being offered a new perspective on the debate.

In this paper, we will first introduce the reviewed approaches and highlight their similarities as well as their advantages (Section 2). Afterwards, we will discuss general challenges caused by differences between the taxonomies and also their individual limitations (Section 3). Subsequently, we will point out ways to overcome these challenges, among others, by presenting an own taxonomy (Section 4). Finally, a discussion about our findings and proposals concludes the paper (Section 5).

2 CURRENT TAXONOMIES OF EXPLAINABILITY METHODS

To get an overview of currently available taxonomies, we reviewed eleven papers from the last three years (2019–2021) referencing, containing, or proposing taxonomies of explainability methods: [6, 10, 14, 30, 31, 44, 49, 54, 63, 65, 72]. While this is by no means a systematic review, we focused on representative papers in the field. Furthermore, we consider the points we will make to be sufficiently general to be applicable to taxonomies that have not been considered. That being said, let us first highlight the commonalities of the reviewed papers by looking at similar distinctions they make.

2.1 Common Distinctions

First of all, there is a distinction between directly training explainable models and explaining a (plausibly opaque) model after it was trained. The former is sometimes called *transparent model (design)* and sometimes *ante-hoc explainability*;[2] the latter uniformly *post-hoc explainability*. Among others, linear regression models, decision trees, k-nearest neighbor models, rule-based learners, general additive models, and Bayesian learners are commonly seen as ante-hoc explainable [10, 14], given they are not too large. What precludes such models from being used for all ML problems is their lack of performance: these models do not commonly achieve an accuracy comparable to that of opaque models, such as DNNs. This is well-known as the performance-explainability trade-off [6, 10, 65].[3]

[1] It should be noted that taxonomies with such discrepancies do anything but live up to their purpose if they are conceived in a strict sense, for example, as ontologies. However, we adopt a more lenient position in this paper, consistent with discussions on explanatory pluralism in the philosophy of science (see, e.g., [28, 33, 36, 48]).

[2] In what follows, we will prefer the term *ante-hoc* explainability over its rival. The reasons for this choice will become visible later on (in Section 3.1.1).

[3] It should be noted that not everyone believes that there is such a trade-off. Rudin [62] argues that most, if not all, ML problems can be satisfyingly solved with ante-hoc methods, given sufficient time and expertise. For this reason, she makes a case for more research into ante-hoc explainable models, especially for high stakes applications.

Since it is commonly assumed that ante-hoc explainable models do not achieve satisfying performance, opaque models are frequently used. These models are so complex that they are black boxes for humans, even eluding the understanding of experts [18]. Moreover, purportedly ante-hoc explainable models suffer from another downside: only small models may retain their understandability. In cases of models with many rules or parameters, these may also become black boxes. In this regard, models that are, in principle, ante-hoc explainable, may need post-hoc methods [44].

This leads us to post-hoc methods. These methods try to generate explanations of already trained models. For post-hoc explainability methods, there is often a further distinction between *model-agnostic* methods and *model-specific* ones. This distinction is about whether the method works for all types of models (model-agnostic) or only specific ones, such as DNNs, support-vector machines (SVMs), or random forests (model-specific).

Furthermore, many papers mention a distinction between explaining a model *locally* (i.e., a single prediction) or *globally* (i.e., the whole model). While this distinction is often made just for post-hoc explainability methods, one could also argue that it makes sense for ante-hoc explainability methods. A very large decision tree might be hard to comprehend as a whole, but an individual classification it makes could be traceable nevertheless.

These distinctions are, as far as we are aware, the only overarching commonalities of the reviewed papers. However, when we examined them individually regarding the way they constructed taxonomies, we found some further similarities in smaller groups of papers. While we were able to create two groups of papers in this way, two further papers defied classification into either of these groups. Both these papers presented sufficiently unique taxonomies of explainability methods to warrant independent classification. In total, then, we formed four categories of papers that differ from each other in the approach they use to construct taxonomies.

2.2 Approaches to Constructing Taxonomies of Explainability Methods

In particular, we call the four approaches to constructing taxonomies we have identified the *functioning-based approach*, the *result-based approach*, the *conceptual approach* and the *mixed approach*. We will examine these approaches in turn.

2.2.1 The Functioning-Based Approach. This approach to constructing taxonomies of explainability methods takes the underlying *functioning* of an explainability method as the essential constituent for its classification. We will go into more detail about what this means in a moment. The taxonomy proposed by Samek and Müller [63] adheres in large parts to this approach, and we will use it for illustrative purposes in what follows. Overall, we find five categories in the functioning-based approach, three of which are taken from Samek and Müller (see Figure 1 for a general visualization).

To describe more precisely what we mean by "functioning" is difficult. In principle, functioning is the way an explainability method extracts information from an ML model. Let us give some examples for further clarification. One classification category proposed by Samek and Müller is *explaining with local perturbations*. Explainability methods that belong to this category perturb a model's inputs slightly in order to find out the *importance* of certain *features* that

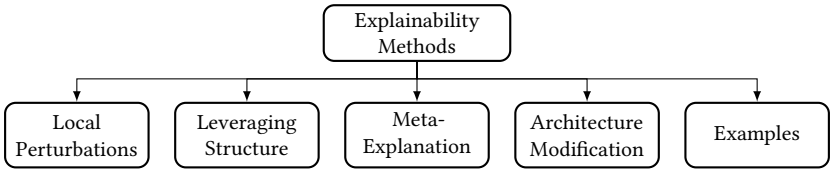

Figure 1: Overarching taxonomy of the *functioning approach*.
From the left, the first three categories are taken from Samek and Müller [63], whereas the other two are taken from Arrieta et al. [10].

this input has on the model's prediction. Accordingly, the functioning behind these methods is *(local) perturbation*.

Samek and Müller also introduce the category *leveraging structure*. The methods in this classification category exploit specific properties of the ML models they are supposed to explain to construct the explanation. In DNNs, for instance, a popular way to leverage structure is to examine gradients. Since gradients are the multivariate generalization of the derivative, they can provide information about the importance of individual input values. Methods that leverage structure often result in *feature importance* attributions, similar to methods that explain with local perturbations.

Another way to explain is by forming *meta-explanations*. Explainability methods with this manner of functioning do not work on an ML model directly, but on explanations for this model generated by other explainability methods. These explanations are aggregated and then compared, with the aim of providing a better explanation than each of the used methods individually. Yet again, these methods often result in *feature importance* attributions.[4]

While *explaining with local perturbations*, *leveraging structure*, and *meta-explanations* are the only functioning-based categories Samek and Müller introduce,[5] there are other interesting ones. For example, Arrieta et al. [10] introduce *architecture modification* as a functioning principle of certain explainability methods. Methods in this category try to simplify complex models by altering their architecture. In convolutional neural networks, for instance, this could be done by exchanging convolutional layers with max-pooling layers (see [67]) [10]. The use of architecture modification can improve the explanations produced by other explainability methods, if not lead to ante-hoc explainable models.

The final category of the functioning-based approach, also introduced by Arrieta et al., is extracting *examples* (see Section 2.2.2).

In our opinion, the functioning-based approach to constructing taxonomies of explainability methods is particularly useful for people interested in developing explainability methods. These people are likely interested in acquiring a picture of different functioning principles at play in various explainability methods. Accordingly, with categories that directly capture the functioning behind such methods, these people will be able to quickly get a rough idea of how most explainability methods function. Based on these insights, they will find it easier to start developing own methods.

[4]One could argue that *meta-explanations* are not way of functioning but rather a product. However, as the product of forming meta-explanations is, as just described, often a feature importance attribution, we list meta-explanations as a *functioning-based approach* to enable a better distinction. This decision is also supported by the linguistic ambiguity that words with the suffix "-ion" can describe a product *and* a process [61].
[5]Samek and Müller additionally introduce *explaining with surrogates*. However, as this is a product of, for instance, explaining with local perturbations, we will not speak further about it at this point. We will come back to this when looking at the *result-based approach* (Section 2.2.2).

2.2.2 The Result-Based Approach. This approach takes the *result* of an explainability method as the essential constituent for its classification. The taxonomy proposed by McDermid et al. [49] adheres mostly to this approach, which has three categories (see Figure 2).

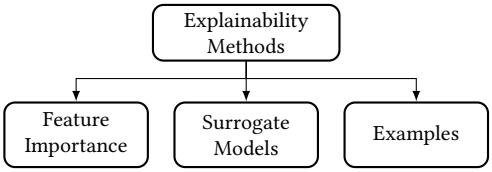

Figure 2: Overarching taxonomy of the *result approach*.

As became evident while discussing the previous approach, many explainability methods aim at uncovering the importance of input features for an output. Accordingly, *feature importance* is a category used by papers adhering to the result-based approach.

Another classification category of the result-based approach is *surrogate models*. Explainability methods that build surrogate models try to approximate (a specific part of) the original model with a simpler, ante-hoc explainable one. Surrogate models can be created in many ways, for instance, by probing the original model via *local perturbations*, or by *leveraging* its *structure*. Accordingly, surrogate models can be the result of most ways of functioning.

A further result with which one can explain is by presenting representative *examples*. For instance, data units from a model's training set that are generating a particularly high (or low) certainty to belong to a specific class can serve as such examples [38]. While examples can be generated or extracted by leveraging a model's structure, there are many further ways to do so that do not correspond to one of the introduced ways of functioning. For this reason, we also introduced this category for functioning-based taxonomies.

The result-based approach to constructing taxonomies of explainability methods might be especially useful for people trying to decide which explainability method to use in an application. These people usually have a good idea of what kind of explanation the application's users need. For instance, examples and feature importance attributions are something that, in many cases, one can easily and quickly understand without much background in ML. Surrogate models, however, are rather something for people with a solid background in ML, who know how to probe such systems.

2.2.3 The Conceptual Approach. This approach splits up the classification of explainability methods into several distinct *conceptual dimensions* that sometimes have hierarchical levels (see Figure 3).

We have already described the core dimensions that this approach uses to classify explainability methods. In particular, these are *stage* (ante-hoc vs. post-hoc), *applicability* (model-agnostic vs.

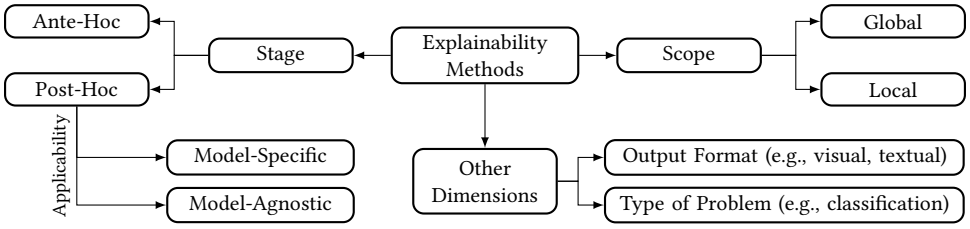

Figure 3: Overarching taxonomy of the *conceptual approach*.
Note that the depicted clusters are not exhaustive, since, e.g, *cohort* has been proposed as a middle-ground between *local* and *global* [65]. Likewise, *model class-specific* has been proposed as a middle-ground between *model-specific* and *model-agnostic* [65].

model-specific), and *scope* (local vs. global).[6] In general, dimensions are mostly independent of each other (except for *applicability*, which only applies to post-hoc methods), though some can be combined more easily with each other (e.g., a global scope makes more sense for ante-hoc methods). The taxonomies proposed in [30, 31, 44, 65, 72] adhere to this approach.

In addition to these core dimensions, different authors have proposed further dimensions. First of all, Sokol and Flach [65] propose more than 30 dimensions of interest when it comes to specifying and classifying explainability methods. Their dimensions also contain more fine-grained categories, introducing levels in between local and global (i.e., *cohort*) as well as in between model-specific and model-agnostic (i.e., *model-class specific*).

Vilone and Longo [72] also discuss more dimensions than stage, applicability, and scope. In particular, their focus lies on *output format*. Here, they distinguish *numerical*, *rules*, *textual*, *visual*, and *mixed*. According to them, this dimension has not received sufficient attention in the literature despite being essential for communicating the result of an explainability method: some output formats might be better suited for certain stakeholders than for others.

Another noteworthy dimension, used by Vilone and Longo as well as Sokol and Flach, is *type of problem*. As the name implies, this dimension concerns for which type(s) of problem the method works (e.g., classification, regression).

The conceptual approach is probably best suited for people who want to get a solid overview of the XAI landscape as a whole, for instance, to work on XAI topics themselves. In particular, the core dimensions give a reliable idea of what matters in the debate and illustrate the most important technical aspects and distinctions of explainability methods in a way that is quickly understood. Additionally, the plug-and-play nature of the conceptual approach, which allows for adding dimensions as desired, is especially well-suited for creating taxonomies that are fit for specific purposes. For these reasons, this approach is likely a good starting point when it comes to (more advanced) interdisciplinary work [17, 41, 44].

2.2.4 The Mixed Approach. Finally, this approach is a hybrid of the above. In other words, papers adhering to this approach use elements of the other three approaches to constructing taxonomies when classifying explainability methods. In particular, the upper levels of the taxonomy are constituted by the stage (ante-hoc vs. post-hoc) and the applicability (model-specific vs. model-agnostic) distinction, as also used in the conceptual approach. On the final level, elements of all previous approaches to constructing taxonomies of explainability methods come into play (see Figure 4). The taxonomies proposed in [6, 10, 14, 54] adhere to this approach, which we will discuss in more detail below.

On the lowest level, all of the reviewed papers in this category follow roughly the same quadripartition: *explanation by simplification*, *explanation by feature relevance*, *visual explanation*, and *local explanation*. While we have already discussed *explanation by feature relevance* and *local explanation*, *visual explanation* and *explanation by simplification* still lack an elaboration. As the name implies, the category *visual explanation* is for explainability methods that visualize their results, for instance, in heat maps. Finally, *explanation by simplification* encompasses functioning principles that result in surrogate models.[7] Despite having this quadripartition in common, there are some differences.

While Arrieta et al. [10] and Minh et al. [54] see all these four categories represented in both model-agnostic and model-specific methods, Angelov et al. [6] as well as Belle and Papantonis [14] exclude visual and local explanations from being model-specific. In contrast, Arrieta and colleagues find even more categories into which model-specific explainability methods can be sorted, for instance, *architecture modification* and *example-based explanations*.

When going further into the details, more differences become visible. While Angelov et al. [6] use the above quadripartition to visualize their taxonomy, what they use to discuss several explainability methods are slightly different categories (see Table 1 or Section 3.1.2). Furthermore, while Belle and Papantonis [14] also visualize their taxonomy by means of this quadripartition, what they discuss is slightly different. On the one hand, they do not give any direct examples for *local explanations*. On the other hand, they list some instances of *example-based explanations*.

The mixed approach has its strength in combining the most important distinctions when it comes to explainability methods. For this reason, this approach is, in our eyes, best suited for people new to XAI. This conclusion is further backed by the fact that the papers employing the mixed approach present a large number of methods (especially [10, 54]) and are, for this reason alone, able to convey a very comprehensive picture of the XAI landscape.

[6] This is not to say that these dimensions do not play a role in the other approaches to constructing taxonomies. In contrast, as we have highlighted above, most reviewed papers make these distinctions in some form. However, what distinguishes the conceptual approach to constructing taxonomies of explainability methods from the other ones is the centrality these dimensions play in forming categories for the classification of explainability methods.

[7] We have deliberately not incorporated this way of functioning above, as the simplification can happen in many ways. Among others, it can happen by using feature perturbation or by leveraging the structure. With this in mind, *simplification* is too much of a general term to be of practical use: more specific terms are available.

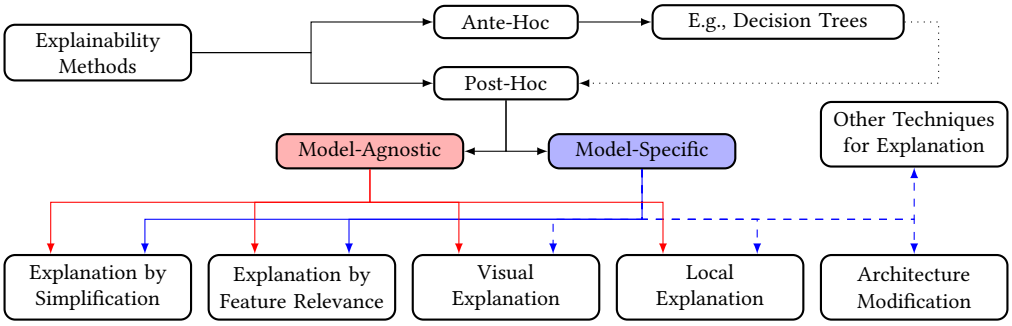

Figure 4: Overarching taxonomy of the *mixed approach*.
Belle and Papantonis [14] admit that purportedly ante-hoc explainable models may require post-hoc explainability methods in some cases (dotted lines). Arrieta et al. [10] and Minh et al. [54] also find *local explanation*, *visual explanation*, *architecture explanation*, and other explanation techniques (e.g., *explanation by example*) as possibilities for model-specific explainability methods (dashed lines).

3 CHALLENGES CONCERNING CURRENT TAXONOMIES

Having described the four approaches to constructing taxonomies that we have identified in the literature and their advantages, we will now discuss challenges these approaches face. In particular, we will first discuss general challenges that arise from the fact that there are several approaches to constructing taxonomies of explainability methods. Subsequently, we will shortly elaborate on individual limitations of the four approaches.

3.1 General Challenges

There are some general challenges that the reviewed taxonomies of explainability methods face. In particular, solely the fact that we could identify four different approaches to constructing such taxonomies can be seen as a challenge, as this may hamper newcomers from acquiring a reliable picture of the XAI landscape.

3.1.1 Misleading Nomenclature. Another factor that may hamper newcomers from easily and efficiently pursuing their goals in XAI is a partially misleading nomenclature. This is a problem that starts even outside of taxonomies of explainability methods, for instance, with the lack of a clear distinction between terms like "explainability" and "interpretability" in the XAI debate [23, 68]. However, it also propagates inside these taxonomies (see, for instance, Table 1).

While it may be unfortunate that different authors speak of *feature-oriented methods*, *feature relevance explanation*, and *feature importance methods*, a competent speaker of the English language may easily see the connection between these terms. It becomes more complicated when authors speak of *explanation by simplification* when alluding to *explaining with surrogate models*. Simplification can mean many things, and one might first think of the original model being modified, for example, as done by explainability methods belonging to the *architecture modification* category.

Another case of misleading nomenclature is, in our eyes, the commonly used synonym of *ante-hoc explainability*: *transparent model design*. What is problematic about this name is that transparency is often equated with understandability in the literature. However, just because something is transparent, this does not mean that it is understandable [5]. For instance, a DNN for which one has access to all weights can be considered transparent, whereas it is most likely not remotely understandable. Particularly scholars from outside of computer science may struggle with such terminology.

3.1.2 No Consensus on Important Categories. Another challenge is the lacking consensus on important categories in these taxonomies. While Table 1 lists the categories that are roughly comparable, there are many more categories that can be found in single taxonomies. In addition to the listed categories, Angelov et al. [6] also propose *global methods*, *concept models*, and *human-centric methods*. Furthermore, Arrieta et al. [10] propose *textual explanations* and *architecture modification*. Finally, Samek and Müller [63] propose *leveraging structure* and *meta-explanation*.

When it comes to the conceptual approach, a greater consensus can be found, though there are other shortcomings. While there is some agreement on the core dimensions (i.e., stage, applicability, and scope), there are many more dimensions that seem plausible. As mentioned above, Sokol and Flach [65] propose more than 30 dimensions that can be used to classify explainability methods. While this may offer a comprehensive picture, such a huge number of dimensions leads to other problems. First and foremost, the overview character that taxonomies normally strive for is lost.[8]

Granted, the differences in classification categories are mainly attributable to the differences in the focus of the individual taxonomies. For this reason, it is only natural that different taxonomies use different categories when classifying explainability methods. However, one would expect differences of this magnitude when it comes to the details of the taxonomies rather than concerning the most important categories. Furthermore, even if these differences are justified, they make it difficult for a person without sufficient background knowledge of XAI to get a good overview.

3.1.3 Differences in Classifying Explainability Methods. The last general challenge we want to mention is that, in some cases, taxonomies differ concerning the classification of one explainability method. This challenge is intensified by the fact that these differences in classification are not only due to differences in classification categories. Even for taxonomies having comparable categories, the classification of one method is sometimes different (see Table 2).

[8] It should be noted that Sokol and Flach [65] do not directly attempt to offer a taxonomy of explainability methods but rather a comprehensive list of attributes that an explainability method can have. However, the main challenge remains: it is not easy to keep track of all these attributes.

Angelov et al. (2021)	Arrieta et al. (2020)	Belle & Papantonis (2021)	McDermid et al. (2021)	Samek & Müller (2019)
Feature-Oriented Methods	Feature Relevance Explanation	Feature Relevance Explanation	Feature Importance Methods	E.g., Explaining With Local Perturbations
Surrogate Models	Explanation by Simplification	Explanation by Simplification	/	Explaining with Surrogates
[Local] Pixel-Based Methods	Visual Explanation	Visual Explanation	/	/
/	Explanation by Example	(Explanation by Example)	Example-Based Methods	/
Local [Pixel-Based] Methods	Local Explanations	(Local Explanations)	/	/

Table 1: Similar categories of different taxonomies compared.
The table contains only categories that can be found across several papers. Note that Minh et al. [54] is not listed in this table, as their taxonomy is, for most parts, congruent to that of Arietta et al. [10]. For the two categories in round brackets, it is not completely clear whether they count as categories in the corresponding paper (see Section 2.2.4). Square brackets indicate that only part of the category is of interest (namely, the part that is *not* in brackets).

Two good examples of explainability methods that are classified in very different manners are *Counterfactual Explanations* [73] and *Testing with Concept Activation Vectors* (TCAV) [39]. For these two methods, the classifications in the reviewed papers do not coincide in the slightest. In particular, counterfactual explanations are classified as a *local explanation* but also as an *example-based explanation* (see first row of Table 2). Likewise, TCAV is classified as a *concept model* but also as a *meta-explanation* (see last row of Table 2).

When it comes to more fine-grained details, the classification of *Integrated Gradients* [69] and *Deep Learning Important FeaTures* (DeepLIFT) [64] has some interesting differences. In addition to the result, McDermid et al. [49] also mention the functioning of explainability methods in the feature importance category. Here, they distinguish *perturbation-based* and *gradient-based* ones. Interestingly, however, they classify the mentioned two methods in a different way than Samek and Müller [63]. According to McDermid and colleagues, DeepLIFT is perturbation-based, whereas integrated gradients is not. However, when looking at the classification of Samek and Müller, the relationship is reversed.

Coming back to our previous finding that it might be unfortunate to use both *explanation by simplification* and *explaining with surrogates* to refer to the same category, here we can find a good example illustrating the underlying problem. Without this knowledge, one of the most famous explainability methods, *Local Interpretable Model-Agnostic Explanation* (LIME) [60], looks as though it is classified in three different ways: *surrogate model*, *explanation by simplification*, and *feature importance*. Especially newcomers to the field of XAI might be confused by this.

3.2 Individual Challenges

Having described the general challenges that the reviewed taxonomies of explainability methods face, let us come to discussing the individual ones. To this end, we will examine each of the four approaches in turn in what follows.

3.2.1 Functioning-Based and Result-Based Approach. We address the challenges of these two approaches simultaneously because they are closely related. Both the functioning-based and the result-based approaches to constructing taxonomies of explainability methods, when taken individually, only draw a limited picture of the XAI landscape. For instance, perturbation-based explainability methods can result in feature visualizations but also in surrogate models. Surrogate models, however, can not only result from perturbations but also from leveraging a model's structure.

In this line of thought, using only one of these classifications seems to withhold crucial information. Taken together, however, the picture painted by these two approaches is more comprehensive. For this reason, using only one of these approaches is not sufficient to adequately classify a method: both approaches are required.

3.2.2 Conceptual Approach. Especially the core dimensions of the conceptual approach are strongly influenced by technical aspects of explainability methods. While this provides a clear frame of reference for classification, scholars coming from outside of computer science might not be interested in such aspects. Psychologists, for instance, could be interested in how the results of a method might appeal to cognitive mechanisms that facilitate understanding.

Some scholars have already noticed this drawback. With their focus on output format, for instance, Vilone and Longo [72] try to overcome it somewhat. According to them, numeric outputs are better suited for experts than for laypersons. Likewise, other output formats are likely better suited in certain contexts than in others.

Langer et al. [44] go one step further and propose a complete model of the XAI pipeline that factors in dimension like scope and applicability and links them to certain stakeholders and their needs. For example, they believe that local explainability methods are likely of greater interest for people affected by the outputs of AI-based systems than for regulators trying to conceive AI legislation: While individuals may want to know whether they have been subjected to a discriminatory decision, regulators presumably aim to curb discrimination at large. Similar preferences are conceivable for developers, who might be more keen on exploring the internals of a model via model-specific methods than via model-agnostic ones.

Although it stands to reason that these technical dimensions provide a good starting point for interdisciplinary research, exclusively

Explainability Method	Angelov et al. (2021)	Belle & Papantonis (2021)	McDermid et al. (2021)	Samek & Müller (2019)	Vilone & Longo (2021)
Counterfactual Explanations [73]	/	Local Explanation	Example-Based Explanation	/	/
DeepLIFT [64]	/	Feature Relevance	Feature Importance (Perturbation-Based)	Leveraging Structure	Visual
Integrated Gradients [69]	/	Feature Relevance	Feature Importance (Gradient-Based)	Explaining with Local Perturbations	Visual
LIME [60]	Surrogate Model	Explanation by Simplification	Feature Importance (Perturbation-Based)	Explaining with Surrogates	Mixed
LRP [9]	Pixel-Based	/	/	Leveraging Structure	Visual
SHAP [45]	Feature-Oriented	Feature Relevance	Feature Importance (Perturbation-Based)	/	Numerical
TCAV [39]	Concept Models	/	/	Meta-Explanation	Numerical

Table 2: Popular explainability methods classified by the reviewed taxonomies.
Although it does not provide comparable categories, we have included the classifications of the *output format* dimension proposed by Vilone and Longo [72] in the rightmost column to illustrate that the output format of a method does not depend on its result or functioning. For instance, feature relevance methods do not necessarily have a visual output format. Furthermore, note that while counterfactual explanations is classified as *local explanation* in the overviews of Belle and Papantonis [14], it is described in the section *explanation by simplification*.

focusing on them might not be sufficient for all the interests that researchers from different disciplines could have. For this reason, more psychologically motivated dimensions might prove valuable.

As a first approximation to such dimensions, the categories of the result-based approach to constructing taxonomies might be helpful. For instance, it was shown that people associate new data with previously learned and aggregated prototypes (i.e., examples, see [15]) [6]. For this reason, using *examples* to explain ML models seems to be psychologically backed. Likewise, feature importance attributions are also valuable from a psychological point of view, as humans are naturally interested in seeing what makes a difference, for instance, by constructing contrast cases [32, 50–52].

3.2.3 Mixed Approach. Overall, taxonomies of the mixed approach should be treated with caution. In the reviewed papers pursuing this approach, the quadripartition is visually situated on one level (as also depicted in Figure 4). This suggests that the categories on this level are of similar importance. Furthermore, it also suggests that these categories are mutually exclusive, a circumstance that is often not explicitly discussed (and, thus, not explicitly excluded) in the respective papers.[9] However, both these allusions are highly misleading: there are explainability methods that, plausibly, can be counted into all four categories (e.g., LIME; see also Table 2).

Moreover, some of these taxonomies exclude visual and local explanations from model-specific explainability methods. However, there is no basis for doing so. Model-specific explainability methods can just focus on one output (i.e., be local). For example, gradient-based methods often do exactly this. Furthermore, such methods can also visualize their explanations (e.g., in heat maps). For this reason, such an exclusion is unfounded and highly misleading. In particular, Arrieta et al. [10] even give examples of model-specific methods that produce local or visual explanations (or both).

[9] A notable exception is Minh et al. [54] who explicitly admit that some explainability methods can be sorted into multiple categories.

3.3 Putting the Challenges Into Context

What becomes clear from these observations is that each of the reviewed taxonomies, taken individually, might not suffice to convey an adequate picture of the XAI landscape. Accordingly, a newcomer to the field of XAI might have insufficient knowledge after studying one taxonomy, as only a combination and a cross-comparison of several taxonomies may lead to an adequate degree of insight. However, such a combination and cross-comparison is made difficult by other challenges we spoke about, such as the missing consensus in nomenclature or classification.

These challenges are mostly not due to the taxonomies. The renewed interest in XAI is still very young, and the field as such has grown to such a degree in recent years that uniformity is barely achievable. Against this background, the task now is to devise ways to overcome the above challenges to achieve more uniformity.

4 THREE WAYS FORWARD

In this section, we will discuss three ways to address the above challenges. First, we will suggest a new taxonomy by combining the discussed approaches to constructing taxonomies. However, some challenges may not be overcome by simply introducing another taxonomy, even if it is more comprehensive. Accordingly, we will discuss two additional ways to overcome the challenges: compiling a database of explainability methods and creating a decision tree to help make decisions about which (type of) explainability method to use. Taken together, these three ways – and the respective artifacts proposed in them – should suffice to overcome many of the above challenges and serve as valuable input for future research.

4.1 Combining the Taxonomies

One way to tackle the individual limitations of the discussed approaches to constructing taxonomies is by synthesizing a new taxonomy that harnesses their individual advantages while avoiding

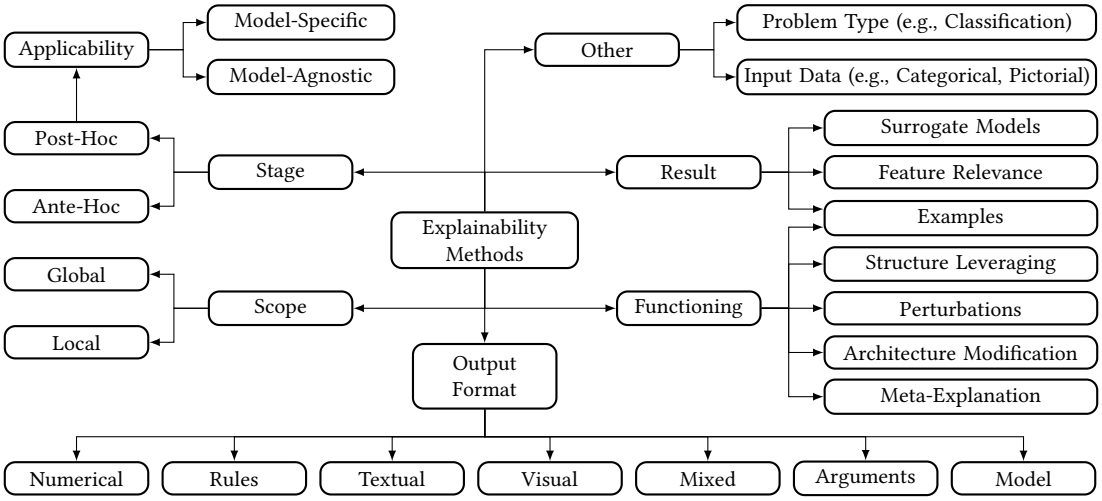

Figure 5: Our suggestion for a taxonomy of explainability methods.

said limitations. A promising starting point for this is to take the conceptual approach as a basis and add two dimensions: *functioning* and *result*, containing the categories of the eponymous approaches. Figure 5 offers a visualization of a taxonomy created in this way.

As already outlined, the plug-and-play nature of the conceptual approach allows for easily adding dimensions. Accordingly, adding the categories of both the *functioning* and the *result-based* approach may provide more psychologically-informed dimensions, while also avoiding the limitations that one of these approaches might be insufficient to provide a comprehensive picture of the XAI landscape. Furthermore, this proposal circumvents the limitations associated with the *mixed approach*, as categories like *local* and *visual* are not situated in the same dimension.

Another noteworthy point is that we do not assume mutually exclusive categories in many dimensions. For instance, the newly added *result* dimension explicitly allows for more than one choice (e.g., LIME creates a surrogate model that highlights the importance of features). Naturally, many dimensions only allow for a single choice in most cases (i.e., an explainability method is in most cases either local or global). However, there are even exceptions to such regularities (see Vilone and Longo [72] for some examples), making the distinctions not as strict as usually proclaimed.

As becomes visible in Figure 5, we have also added the dimension *output format*. We made this addition because output format is, in our eyes, an important ingredient of XAI taxonomies. The reason for this is described above: the output format likely influences the usefulness of an explainability method for certain stakeholders and contexts. For instance, numerical outputs may, as claimed by Vilone and Longo [72], not be suited for laypersons. Likewise, textual outputs require parsing time, plausibly limiting their suitability for situations in which quick decisions must be made.

In addition to the categories proposed by Vilone and Longo, we further opt for the categories *arguments* and *model*. There are many scholars who argue that *arguments* are a useful but unnoticed way to explain decisions (see, e.g., [3, 11, 12, 66]). The idea is that arguments are, in principle, a way that humans use to come to decisions and, thus, particularly well-suited to help a person understand a decision. Additionally, there are computational frameworks designed for modeling arguments, providing a good starting point for further research [3, 26]. In principle, arguments consist in presenting features (visually, textually, or even numerically) that contributed to a decision and features that were detrimental to it. For this reason, arguments may be a good way to present feature relevance.

Adding *models* as an output for explainability methods factors in that some methods produce surrogate models that are not further processed to match one of the other output formats. In particular, explainability methods of the architecture modification category often produce a modified model that is not meant for presentation, but rather for subsequent use by other explainability methods.

Obviously, the proposed taxonomy is not exhaustive and many more dimensions, as well as many more categories in the proposed dimensions, are possible.[10] However, we believe that our proposal has a pragmatically adequate level of detail to be sufficient for most interests in XAI while at the same time not being too overwhelming.

4.2 Compiling a Database of Explainability Methods

The goal of a taxonomy of explainability methods should be to provide an overview of the XAI landscape. However, without examples of methods that are actually classified, such taxonomies remain theoretical artifacts whose imminent practical use remains to be seen. Accordingly, the more methods are actually classified, the more useful the taxonomy actually is. In this line of thought, papers proposing taxonomies often give some examples for illustration (see Table 2 for some examples given in the reviewed papers).

However, it stands to reason that while giving examples may be useful to illustrate a taxonomy, the best use of it is to comprehensively classify as many methods as possible. To this end, we propose that a database of explainability methods, with classifications pertaining to, for instance, our proposed taxonomy, would be a valuable step forward. Let us discuss this proposal briefly.

[10] More dimensions, as well as more categories in them, can be found in [65].

The idea is that the taxonomy serves as an initial orientation help and the database is structured based on the proposed dimensions. Accordingly, individuals with different interests may first identify dimensions that are most conducive to their goals (e.g., psychologists might be more interested in *result* or in *output format* than in other dimensions) to later look up explainability methods for specific purposes (e.g., methods that visualize feature relevance).

With their proposal of *explainability fact sheets*, Sokol and Flach [65] already embarked on creating something like the database we envision. Their idea is to comprehensively classify explainability methods with more than 30 dimensions to facilitate comparison between them. However, the database we envision would go one step further, being freely searchable and sortable at will, while also enabling the cross-comparison of explainability methods.

4.3 Creating a Decision Tree for Choosing Fitting Explainability Methods

The final idea we want to discuss is a decision tree that should help make decisions about which (type of) explainability method to use. This idea is inspired by the work of Arya et al. [8], who built a taxonomy based on questions that developers might ask themselves to find out which method to take for the purpose at hand. Likewise, one could link the categories of one of the discussed taxonomies to such questions, supporting the usefulness of that taxonomy.

The chosen questions would build up on each other, such that they form a decision tree one traverses in order to find the desired (type of) method. On a high level, a question could be whether the system is already extant and should be supported by an explanation component, or whether a new system should be designed that is already explainable. This question would roughly correspond to the *stage* distinction. Further down, one could find questions that concern the *functioning* or the *result* of the used explainability method. Importantly, similar questions may come up at different points in the decision tree, making room for the possibility of the same (type of) method being useful for different purposes.[11]

Furthermore, there may not be just one such decision tree, but several ones, differing, for instance, based on the area of application. In application areas where human lives are at stake, for instance, some explainability methods might not be suitable, just as Rudin [62] argues (more on this later). Accordingly, decision trees for these areas would look different from those in which there is not much at stake. Additionally, decision trees might not only be helpful for developers. XAI is an increasingly interdisciplinary area of research [44]. However, researchers from outside of computer science might have a hard time starting on the topic. For this reason, decision trees specifically for people from outside of computer science could be created, helping them find explainability methods of interest.

Overall, the three artifacts proposed here (i.e., combined taxonomy, database, decision tree) may prove to be valuable to overcome the above challenges. The combined taxonomy should structure the field of XAI and serve as an initial overview. Furthermore, the database should function as reference work for in-depth knowledge of the field. Finally, the decision tree should guide newcomers and experts alike, building a bridge between taxonomy and database.

[11] In this aspect, we differ from the artifact brought forward by Arya et al. [8], as they strive for a genuine taxonomy that has no question more than once.

5 DISCUSSION

Let us take a step back. While the three artifacts above should be sufficient to cover many interests concerning XAI, they might fail to do so in certain cases. The complexity of the field of XAI, as well as the diverse interests researchers from different disciplines have when engaging in XAI will likely lead to cases that are too idiosyncratic to be solvable by one of our proposals. In this section we will shortly discuss some ways that may, then, do the trick.

5.1 Taxonomies for Scholars From Outside of Computer Science

As discussed, each of the reviewed taxonomies has its own raison d'être. However, as we have argued, these taxonomies are not suited for all purposes one might pursue with a taxonomy. For instance, we have used the criterion of newcomer-friendliness a few times when judging the quality of these taxonomies. Here one could argue that many of these taxonomies are not meant for newcomers to the field of XAI, at least not for ones from outside of computer science. Even for our proposal, we have to admit that it is only a first step to facilitate interdisciplinary research. Taking this into account, it could be valuable to devise taxonomies that are tailor-made for scholars from outside of computer science.

One idea in this direction can be found in Tim Miller's seminal work "Explanation in Artificial Intelligence: Insights From the Social Sciences" [51]. For this work, he reviewed the literature on explanation from philosophy, cognitive science, and psychology to inform computer scientists about research on explanations. What motivated him is the observation that the XAI literature has no agreement on what makes for a good explanation: together with colleagues, he discovered that many authors just propose what they feel to be good explanations without any justification [53].

In his review [51], Miller found four properties that explanations should fulfill: they should be contrastive, selected for the purpose at hand, not contain probabilities, and respect social aspects (e.g., factor in the background knowledge of a person). Furthermore, he found that different types of explanation each have their own advantages and drawbacks. When it comes to evoking understanding, for instance, some are better than others. Hence, a taxonomy of explainability methods that is based on the type of produced explanation might be helpful.

In particular, philosophers and psychologists are professionally concerned with different types of explanation. Accordingly, such an explanation-based taxonomy could help people from these disciplines to get started with XAI. For instance, utilizing such a taxonomy, psychologists could conduct studies to find out which kinds of explanations are best suited for certain contexts (see, e.g., [42, 43, 71] for such kinds of studies in hiring scenarios).

5.2 Taxonomies and Reviews Focusing on Specific Aspects

Devising taxonomies for narrow application areas could also be a way to address missing cases. Indeed, there are already some works that go into this direction. For instance, Müller et al. [55] construct a taxonomy solely for the domain of computational pathology, and Tjoa and Guan [70] construct one for the medical domain.

Other papers do not present taxonomies but overviews of explainability for specific domains. For example, Nunes and Jannach [56] focus on explainability of recommender systems, Anjomshoae et al. [7] on explainability of robots and human-robot interaction, Abdul et al. [1] on the human-computer interaction (HCI) domain, and Mathews [47] on biomedical and malware classification.

Likewise, reviews that focus on one category of explainability methods could also prove to be useful in order to cover cases for which our proposals might fail. Ivanovs et al. [35], for instance, review perturbation-based explainability methods for DNNs, finding and highlighting differences as well as similarities between them.

5.3 New Dimensions for the Taxonomy

Another factor that might contribute to the emergence of cases that our proposals do not cover is that vital dimensions are missing. We will illustrate this idea by means of the dimension *fidelity* [41].

Many post-hoc explainability methods suffer from the fact that the explanation they generate does not necessarily represent what is going on in the model they are explaining [2, 4, 39]. This prompted Cynthia Rudin to write a paper in which she pleads for an end to the use of opaque models in high-stakes situations [62]: all we can do to explain such models is to apply post-hoc explainability methods that suffer from the problem just described. Accordingly, it is hard to find out whether the decision was based on valid criteria, and it is also hard to allocate responsibility for it.

However, this problem mostly concerns model-agnostic explainability methods, as these methods do not factor in the internals of the models they explain. Model-specific explainability methods fare better in this regard: they exploit a model's internals to construct explanations. Nevertheless, there is still a continuum of what degree of fidelity to the explained models the explanations have. Accordingly, one could taxonomize methods based on their fidelity.

One motivation for such a dimension is as follows: different scenario have different stakes. The higher the stakes, the more important it becomes to be able to receive a fidelitous explanation in case of failure. In some cases, there might be no way around using ante-hoc explainability methods, just as Rudin argues [62].

5.4 New Levels for the Taxonomy

The last idea we want to discuss in order to address cases that our artifacts may not be able to solve is adding new levels to the taxonomy. We will discuss this for the *result* dimension.

To this end, let us return to the methods *counterfactual explanations* [73] and *TCAV* [39], which were previously classified incongruously. Our taxonomy can, in principle, do them justice. When it comes to the result, both belong to the *feature relevance* category.

However, the features whose relevance they quantify are semantically quite different. Counterfactual explanations indicate what change in input values would be required to achieve a particular output value. TCAV, on the other hand, indicates the relevance of certain user-defined concepts to a particular classification class.

While some of these semantic differences could be due to the difference in scope (counterfactual explanations are local, while TCAV is global), there is still a difference between using available features (given by the input variables) and using newly constructed ones (as defined by the user). Therefore, to capture this difference, one could introduce new subcategories for feature relevance methods. These subcategories could pertain to the (type of) feature examined.

In fact, some of the examined taxonomies have such subcategories (e.g., [10, 14]). However, the incongruity is even more pronounced for these categories than for the more general ones in the taxonomies. Thus, there is still much to be explored in this regard.

6 CONCLUSION AND FUTURE WORK

Research in the field of XAI is on the rise. In this paper, we have reviewed several recent approaches to constructing taxonomies of explainability methods. Among other things, we have pointed out that the differences and inconsistencies between these taxonomies may not allow us to obtain a clear picture of the XAI landscape smoothly. For XAI to unlock its full potential, however, such a clear picture is indispensable.

The problems that XAI should tackle often require interdisciplinary research [17, 41, 44]. Preventing discrimination [24, 41, 74], increasing trustworthiness [37, 41, 46], allocating responsibility [13, 41, 59], and generally, promoting human well-being [27, 46] is, in principle, possible with explainability – as long as researchers from different disciplines can come together to work on it. Accordingly, confusion in the field may postpone the potentially vast social benefits XAI promises to bring about.

Based on our analysis of the research field and its problems, we proposed and discussed three artifacts that seem promising to provide a remedy: a taxonomy combining elements of the reviewed ones, a database of explainability methods, and a decision tree to help decide which (type of) method is needed. We believe that future research into these three artifacts will provide a valuable basis for other projects in the realm of XAI, especially interdisciplinary ones.

ACKNOWLEDGMENTS

Work on this paper was funded by the Volkswagen Foundation grant AZ 98514 "Explainable Intelligent Systems" (EIS) and by the DFG grant 389792660 as part of TRR 248. The Volkswagen Foundation and the DFG had no role in preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication. The authors declare no other financial interests.

We would like to thank all participants of the EIS reading group for initial discussions about the topic and feedback on an early version of this article. Furthermore, we would like to thank three anonymous reviewers and all participants of the EIS colloquium for helpful comments on a more advanced version of this article.

Special thanks go to Sarah Sterz for motivating us to write the article, Sara Mann and Markus Langer for detailed comments, and Barnaby Crook for proofreading. All remaining errors are our own.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 Conference on Human Factors in Computing Systems* (Montréal, Québec, Canada) *(CHI 2018)*, Regan L. Mandryk, Mark Hancock, Mark Perry, and Anna L. Cox (Eds.). Association for Computing Machinery, New York, NY, USA, Article 582, 18 pages. https://doi.org/10.1145/3173574.3174156

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2020. Sanity Checks for Saliency Maps. arXiv:1810.03292

[3] Leila Amgoud and Henri Prade. 2009. Using Arguments for Making and Explaining Decisions. *Artificial Intelligence* 173, 3–4 (2009), 413–436. https:

//doi.org/10.1016/j.artint.2008.11.006

[4] Elvio G. Amparore, Alan Perotti, and Paolo Bajardi. 2021. To Trust or Not to Trust an Explanation: Using LEAF to Evaluate Local Linear XAI Methods. *PeerJ Computer Science* 7 (2021), 1–26. https://doi.org/10.7717/peerj-cs.479

[5] Mike Ananny and Kate Crawford. 2018. Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability. *New Media & Society* 20, 3 (2018), 973–989. https://doi.org/10.1177/1461444816676645

[6] Plamen P. Angelov, Eduardo A. Soares, Richard M. Jiang, Nicholas I. Arnold, and Peter M. Atkinson. 2021. Explainable Artificial Intelligence: An Analytical Review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 11, 5, Article e1424 (2021), 13 pages. https://doi.org/10.1002/widm.1424

[7] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable Agents and Robots: Results from a Systematic Literature Review. In *Proceedings of the 18th International Conference on Autonomous Agents and Multi-Agent Systems* (Montréal, Québec, Canada) *(AAMAS 2019)*, Edith Elkind, Manuela Veloso, Noa Agmon, and Matthew E. Taylor (Eds.). International Foundation for Autonomous Agents and Multiagent Systems, Richland County, SC, USA, 1078–1088. https://doi.org/10.5555/3306127.3331806

[8] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. 2021. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. arXiv:1909.03012

[9] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS ONE* 10, 7 (2015), 1–46. https://doi.org/10.1371/journal.pone.0130140

[10] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Benjamins Richard, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[11] Kevin Baum, Holger Hermanns, and Timo Speith. 2018. From Machine Ethics to Machine Explainability and Back. In *International Symposium on Artificial Intelligence and Mathematics* (Fort Lauderdale, Florida, USA) *(ISAIM 2018)*. International Symposium on Artificial Intelligence and Mathematics, Fort Lauderdale, FL, USA, 1–8. https://isaim2018.cs.ou.edu/papers/ISAIM2018_Ethics_Baum_etal.pdf

[12] Kevin Baum, Holger Hermanns, and Timo Speith. 2018. Towards a Framework Combining Machine Ethics and Machine Explainability. In *Proceedings of the 3rd Workshop on Formal Reasoning about Causation, Responsibility, and Explanations in Science and Technology* (Thessaloniki, Greece) *(CREST 2018)*, Bernd Finkbeiner and Samantha Kleinberg (Eds.). Electronic Proceedings in Theoretical Computer Science, Sydney, AU, 34–49. https://doi.org/10.4204/EPTCS.286.4

[13] Kevin Baum, Susanne Mantel, Eva Schmidt, and Timo Speith. 2022. From Responsibility to Reason-Giving Explainable Artificial Intelligence. *Philosophy & Technology* 35, 1 (2022), 1–30. https://doi.org/10.1007/s13347-022-00510-w

[14] Vaishak Belle and Ioannis Papantonis. 2021. Principles and Practice of Explainable Machine Learning. *Frontiers in Big Data* 4, Article 688969 (2021), 25 pages. https://doi.org/10.3389/fdata.2021.688969

[15] Jacob Bien and Robert Tibshirani. 2011. Prototype Selection for Interpretable Classification. *The Annals of Applied Statistics* 5, 4 (2011), 2403–2424. https://doi.org/10.1214/11-AOAS495

[16] David C. Brock. 2018. Learning from Artificial Intelligence's Previous Awakenings: The History of Expert Systems. *AI Magazine* 39, 3 (2018), 3–15. https://doi.org/10.1609/aimag.v39i3.2809

[17] Wasja Brunotte, Larissa Chazette, Verena Klös, and Timo Speith. 2022. Quo Vadis, Explainability? – A Research Roadmap for Explainability Engineering. In *Requirements Engineering: Foundation for Software Quality*, Vincenzo Gervasi and Andreas Vogelsang (Eds.). Springer International Publishing, Cham, CH, 26–32. https://doi.org/10.1007/978-3-030-98464-9_3

[18] Jenna Burrell. 2016. How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms. *Big Data & Society* 3, 1 (2016), 1–12. https://doi.org/10.1177/2053951715622512

[19] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* 8, 8, Article 832 (2019), 34 pages. https://doi.org/10.3390/electronics8080832

[20] Davide Castelvecchi. 2016. Can we open the black box of AI? *Nature* 538, 7623 (2016), 20–23. https://doi.org/10.1038/538020a

[21] Larissa Chazette, Wasja Brunotte, and Timo Speith. 2021. Exploring Explainability: A Definition, a Model, and a Knowledge Catalogue. In *IEEE 29th International Requirements Engineering Conference* (South Bend, Indiana, USA) *(RE 2021)*, Jane Cleland-Huang, Ana Moreira, Kurt Schneider, and Michael Vierhauser (Eds.). IEEE, Piscataway, NJ, USA, 197–208. https://doi.org/10.1109/RE51729.2021.00025

[22] William J. Clancey. 1983. The Epistemology of a Rule-Based Expert System – A Framework for Explanation. *Artificial Intelligence* 20, 3 (1983), 215–251. https://doi.org/10.1016/0004-3702(83)90008-5

[23] Miruna-Adriana Clinciu and Helen Hastie. 2019. A Survey of Explainable AI Terminology. In *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence* (Tokyo, Japan) *(NL4XAI 2019)*, Jose M. Alonso and Alejandro Catala (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, 8–13. https://doi.org/10.18653/v1/W19-8403

[24] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California, USA) *(IUI 2019)*. Association for Computing Machinery, New York, NY, USA, 275–285. https://doi.org/10.1145/3301275.3302310

[25] Julia Dressel and Hany Farid. 2018. The Accuracy, Fairness, and Limits of Predicting Recidivism. *Science Advances* 4, 1 (2018), 1–5. https://doi.org/10.1126/sciadv.aao5580

[26] Phan Minh Dung. 1995. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artificial Intelligence* 77, 2 (1995), 321–357. https://doi.org/10.1016/0004-3702(94)00041-X

[27] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. 2018. AI4People — An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines* 28, 4 (2018), 689–707. https://doi.org/10.1007/s11023-018-9482-5

[28] Victor Gijsbers. 2016. Explanatory Pluralism and the (Dis)Unity of Science: The Argument from Incompatible Counterfactual Consequences. *Frontiers in Psychiatry* 7, Article 32 (2016), 10 pages. https://doi.org/10.3389/fpsyt.2016.00032

[29] Leilani H. Gilpin, Cecilia Testart, Nathaniel Fruchter, and Julius Adebayo. 2019. Explaining Explanations to Society. In *Proceedings of the NeurIPS 2018 Workshop on Ethical, Social and Governance Issues in AI* (Montréal, Québec, Canada), Chloé Bakalar, Sarah Bird, Tiberio Caetano, Edward Felten, Dario Garcia-Garcia, Isabel Kloumann, Finn Lattimore, Sendhil Mullainathan, and D. Sculley (Eds.). 1–6. arXiv:1901.06560

[30] Riccardo Guidotti, Anna Monreale, Dino Pedreschi, and Fosca Giannotti. 2021. Principles of Explainable Artificial Intelligence. In *Explainable AI Within the Digital Transformation and Cyber Physical Systems: XAI Methods and Applications*, Moamar Sayed-Mouchaweh (Ed.). Springer International Publishing, Cham, CH, Chapter 2, 9–31. https://doi.org/10.1007/978-3-030-76409-8_2

[31] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A Survey of Methods for Explaining Black Box Models. *Comput. Surveys* 51, 5 (2019), 1–42. https://doi.org/10.1145/3236009

[32] Denis J Hilton. 1990. Conversational processes and causal explanation. *Psychological Bulletin* 107, 1 (1990), 65–81. https://doi.org/10.1037/0033-2909.107.1.65

[33] Eric Hochstein. 2022. Foregrounding and Backgrounding: A New Interpretation of "Levels" in Science. *European Journal for Philosophy of Science* 12, 2, Article 23 (2022), 22 pages. https://doi.org/10.1007/s13194-022-00457-x

[34] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. 2019. Causability and Explainability of Artificial Intelligence in Medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9, 4, Article e1312 (2019), 13 pages. https://doi.org/10.1002/widm.1312

[35] Maksims Ivanovs, Roberts Kadikis, and Kaspars Ozols. 2021. Perturbation-Based Methods for Explaining Deep Neural Networks: A Survey. *Pattern Recognition Letters* 150 (2021), 228–234. https://doi.org/10.1016/j.patrec.2021.06.030

[36] Lena Kästner. 2018. Integrating Mechanistic Explanations Through Epistemic Perspectives. *Studies in History and Philosophy of Science Part A* 68 (2018), 68–79. https://doi.org/10.1016/j.shpsa.2018.01.011

[37] Lena Kästner, Markus Langer, Veronika Lazar, Astrid Schomäcker, Timo Speith, and Sarah Sterz. 2021. On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness. In *29th IEEE International Requirements Engineering Conference Workshops* (Notre Dame, Indiana, USA) *(REW 2021)*, Tao Yue and Mehdi Mirakhorli (Eds.). IEEE, Piscataway, NJ, USA, 169–175. https://doi.org/10.1109/REW53955.2021.00031

[38] Been Kim, Oluwasanmi Koyejo, and Rajiv Khanna. 2016. Examples are not enough, learn to criticize! Criticism for Interpretability. In *Advances in Neural Information Processing Systems 29* (Barcelona, Spain), Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.). Curran Associates, Inc., New York, NY, USA, 2280–2288. https://proceedings.neurips.cc/paper/2016/hash/5680522b8e2bb01943234bce7bf84534-Abstract.html

[39] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning* (Stockholm, Sweden) *(ICML 2018)*, Jennifer G. Dy and Andreas Krause (Eds.). Microtome Publishing, Brookline, MA, USA, 2668–2677. http://proceedings.mlr.press/v80/kim18d.html

[40] Maximilian A. Köhl, Kevin Baum, Dimitri Bohlender, Markus Langer, Daniel Oster, and Timo Speith. 2019. Explainability as a Non-Functional Requirement. In *IEEE 27th International Requirements Engineering Conference* (Jeju Island, Republic of Korea) *(RE 2019)*, Daniela E. Damian, Anna Perini, and Seok-Won Lee (Eds.).

IEEE, Piscataway, NJ, USA, 363–368. https://doi.org/10.1109/RE.2019.00046

[41] Markus Langer, Kevin Baum, Kathrin Hartmann, Stefan Hessel, Timo Speith, and Jonas Wahl. 2021. Explainability Auditing for Intelligent Systems: A Rationale for Multi-Disciplinary Perspectives. In *29th IEEE International Requirements Engineering Conference Workshops* (Notre Dame, Indiana, USA) *(REW 2021)*, Tao Yue and Mehdi Mirakhorli (Eds.). IEEE, Piscataway, NJ, USA, 164–168. https://doi.org/10.1109/REW53955.2021.00030

[42] Markus Langer, Kevin Baum, Cornelius J. König, Viviane Hähne, Daniel Oster, and Timo Speith. 2021. Spare Me the Details: How the Type of Information About Automated Interviews Influences Applicant Reactions. *International Journal of Selection and Assessment* 29, 2 (2021), 154–169. https://doi.org/10.1111/ijsa.12325

[43] Markus Langer, Cornelius J. König, and Andromachi Fitili. 2018. Information as a Double-Edged Sword: The Role of Computer Experience and Information on Applicant Reactions Towards Novel Technologies for Personnel Selection. *Computers in Human Behavior* 81 (2018), 19–30. https://doi.org/10.1016/j.chb.2017.11.036

[44] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What Do We Want From Explainable Artificial Intelligence (XAI)? – A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research. *Artificial Intelligence* 296, Article 103473 (2021), 24 pages. https://doi.org/10.1016/j.artint.2021.103473

[45] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30* (Long Beach, California, USA) *(NIPS 2017)*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). Curran Associates, Inc., New York, NY, USA, 4765–4774. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[46] Aniek F. Markus, Jan A. Kors, and Peter R. Rijnbeek. 2021. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics* 113, Article 103655 (2021), 11 pages. https://doi.org/10.1016/j.jbi.2020.103655

[47] Sherin Mary Mathews. 2019. Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A Literature Review. In *Intelligent Computing – Proceedings of the Computing Conference* (London, England, United Kingdom) *(CompCom 2019)*, Kohei Arai, Rahul Bhatia, and Supriya Kapoor (Eds.). Springer International Publishing, Cham, CH, 1269–1292. https://doi.org/10.1007/978-3-030-22868-2_90

[48] Robert N. McCauley and William Bechtel. 2001. Explanatory pluralism and heuristic identity theory. *Theory & Psychology* 11, 6 (2001), 736–760. https://doi.org/10.1177/0959354301116002

[49] John A. McDermid, Yan Jia, Zoe Porter, and Ibrahim Habli. 2021. Artificial Intelligence Explainability: The Technical and Ethical Dimensions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379, 2207, Article 20200363 (2021), 18 pages. https://doi.org/10.1098/rsta.2020.0363

[50] Ann L McGill and Jill G Klein. 1993. Contrastive and counterfactual reasoning in causal judgment. *Journal of Personality and Social Psychology* 64, 6 (1993), 897–905. https://doi.org/10.1037/0022-3514.64.6.897

[51] Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[52] Tim Miller. 2021. Contrastive explanation: A structural-model approach. *The Knowledge Engineering Review* 36, Article e14 (2021), 22 pages. https://doi.org/10.1017/S0269888921000102

[53] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of Inmates Running the Asylum. Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. In *Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence* (Melbourne, Australia) *(IJCAI XAI 2017)*, David W. Aha, Trevor Darrell, Michael Pazzani, Darryn Reid, Claude Sammut, and Peter Stone (Eds.). 36–42. arXiv:1712.00547

[54] Dang Minh, H. Xiang Wang, Y. Fen Li, and Tan N. Nguyen. 2021. Explainable Artificial Intelligence: A Comprehensive Review. https://doi.org/10.1007/s10462-021-10088-y Online First in *Artificial Intelligence Review*.

[55] Heimo Müller, Michaela Kargl, Markus Plass, Bettina Kipperer, Luka Brcic, Peter Regitnig, Christian Geißler, Tobias Küster, Norman Zerbe, and Andreas Holzinger. 2022. Towards a Taxonomy for Explainable AI in Computational Pathology. In *Humanity Driven AI: Productivity, Well-being, Sustainability and Partnership*, Fang Chen and Jianlong Zhou (Eds.). Springer International Publishing, Cham, CH, Chapter 15, 311–330. https://doi.org/10.1007/978-3-030-72188-6_15

[56] Ingrid Nunes and Dietmar Jannach. 2017. A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems. *User Modeling and User-Adapted Interaction* 27, 3–5 (2017), 393–444. https://doi.org/10.1007/s11257-017-9195-0

[57] Andrés Páez. 2019. The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds & Machines* 29, 3 (2019), 441–459. https://doi.org/10.1007/s11023-019-09502-w

[58] Arjun Panesar. 2019. Ethics of Intelligence. In *Machine Learning and AI for Healthcare: Big Data for Improved Health Outcomes*, Arjun Panesar (Ed.). Apress, Berkeley, CA, USA, 207–254. https://doi.org/10.1007/978-1-4842-3799-1_6

[59] Wolter Pieters. 2011. Explanation and trust: what to tell the user in security and AI? *Ethics and Information Technology* 13, 1 (2011), 53–64. https://doi.org/10.1007/s10676-010-9253-3

[60] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD 2016)*, Charu Aggarwal, Balaji Krishnapuram, Rajeev Rastogi, Dou Shen, Mohak Shah, and Alex Smola (Eds.). Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[61] David-Hillel Ruben. 2015. *Explaining Explanation*. Routledge, New York, NY, USA. https://doi.org/10.4324/9781315634739

[62] Cynthia Rudin. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[63] Wojciech Samek and Klaus-Robert Müller. 2019. Towards Explainable Artificial Intelligence. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller (Eds.). Springer International Publishing, Cham, CH, Chapter 1, 5–22. https://doi.org/10.1007/978-3-030-28954-6_1

[64] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning Important Features Through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning* (Sydney, Australia) *(ICML 2017)*, Doina Precup and Yee Whye Teh (Eds.). Microtome Publishing, Brookline, MA, USA, 3145–3153. http://proceedings.mlr.press/v70/shrikumar17a.html

[65] Kacper Sokol and Peter Flach. 2020. Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* 2020)*, Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna (Eds.). Association for Computing Machinery, New York, NY, USA, 56–67. https://doi.org/10.1145/3351095.3372870

[66] Kacper Sokol and Peter A. Flach. 2017. The Role of Textualisation and Argumentation in Understanding the Machine Learning Process. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, Carles Sierra (Ed.). IJCAI Organization, Santa Clara, USA, 5211–5212. https://doi.org/10.24963/ijcai.2017/765

[67] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. 2015. Striving for Simplicity: The All Convolutional Net. In *Proceedings of the 3rd International Conference on Learning Representations Workshop Track* (San Diego, California, USA) *(ICLR WT 2015)*, Yoshua Bengio and Yann LeCun (Eds.). 1–14. arXiv:1412.6806

[68] Sarah Sterz, Kevin Baum, Anne Lauber-Rönsberg, and Holger Hermanns. 2021. Towards Perspicuity Requirements. In *29th IEEE International Requirements Engineering Conference Workshops* (Notre Dame, Indiana, USA) *(REW 2021)*, Tao Yue and Mehdi Mirakhorli (Eds.). IEEE, Piscataway, NJ, USA, 159–163. https://doi.org/10.1109/REW53955.2021.00029

[69] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning* (Sydney, New South Wales, Australia) *(ICML 2017)*, Tony Jebara, Doina Precup, and Yee Whye Teh (Eds.). Microtome Publishing, Brookline, MA, USA, 3319–3328. http://proceedings.mlr.press/v70/sundararajan17a.html

[70] Erico Tjoa and Cuntai Guan. 2021. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems* 32, 11 (2021), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314

[71] Donald M. Truxillo, Todd E. Bodner, Marilena Bertolino, Talya N. Bauer, and Clayton A. Yonce. 2009. Effects of Explanations on Applicant Reactions: A meta-analytic review. *International Journal of Selection and Assessment* 17, 4 (2009), 346–361. https://doi.org/10.1111/j.1468-2389.2009.00478.x

[72] Giulia Vilone and Luca Longo. 2021. Classification of Explainable Artificial Intelligence Methods through Their Output Formats. *Machine Learning and Knowledge Extraction* 3, 3 (2021), 615–661. https://doi.org/10.3390/make3030032

[73] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* 31, 2 (2017), 841–887. https://doi.org/10.2139/ssrn.3063289

[74] Jianlong Zhou, Fang Chen, and Andreas Holzinger. 2022. Towards Explainability for AI Fairness. In *xxAI – Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, Andreas Holzinger, Randy Goebel, Ruth Fong, Taesup Moon, Klaus-Robert Müller, and Wojciech Samek (Eds.). Springer International Publishing, Cham, CH, Chapter 18, 375–386. https://doi.org/10.1007/978-3-031-04083-2_18

[75] Jianlong Zhou, Amir H. Gandomi, Fang Chen, and Andreas Holzinger. 2021. Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics* 10, 5, Article 593 (2021), 19 pages. https://doi.org/10.3390/electronics10050593