Approaching Principles of XAI: A SYSTEMATIZATION

Raphael Ronge, Bernhard Bauer, and Benjamin Rathgeber

Abstract—**Today's Explainable Artificial Intelligence (XAI) landscape is the product of a long history of ever-changing Artificial Intelligence (AI) research and attempts to explain it. It can be vast and confusing. Our historical reconstruction of XAI developments relates AI improvements to their inevitable impact on explanation research. The reconstruction provides the basis for an analysis of the state of XAI and for discussing its future developments in our paper and in general. We then propose a new taxonomy based on this historical reconstruction and current XAI approaches. It is a balanced mixture of detail and general applicability. It is therefore intended to be useful in a wide variety of contexts. The flowchart inspired nature of our taxonomy relates its dimensions not only to the XAI development process, but also to each other, creating an additional layer of structure. Given the historical reconstruction and our taxonomy, we are able to propose three principles: *Computing Edges*, *Dimensionality Reduction*, and *Traceability/Blaming*. These are capable of structuring the debate in a new way, as they are not intended to be just ideas that current approaches adhere to. We also propose two new principles for the future (*Embedment* and *Scientific Testing*) that XAI approaches should adhere to in order to improve their explanations. Our findings provide a structured approach to the analysis and development of XAI methodologies. By integrating historical perspectives with state-of-the-art approaches, our research provides a basis for stimulating discussion about the principles that XAI follows and should follow in the future.**

Impact Statement—**The lack of transparency of modern Machine Learning solutions has led to a proliferation of Explainable Artificial Intelligence (XAI) approaches and, by now, various taxonomies of these approaches. This paper is a meaningful addition to these taxonomies, not only by descriptively listing XAI papers, but also by contextualising the state-of-the-art with a historical reconstruction and an analysis of the overarching principles of current research. Through this comprehensive approach, our paper is able to provide a better understanding of the prospects and challenges of XAI and to help systematise new approaches. In doing so, it serves as a basis for discussion of XAI approaches and for future analyses of XAI beyond its technical foundation.**

Index Terms—**Explainable Artificial Intelligence, Interpretable Artificial Intelligence, Interpretable Machine Learning, Explanation**

R. Ronge, B. Bauer and B. Rathgeber, "Approaching Principles of XAI: A Systematization," in IEEE Transactions on Artificial Intelligence, doi: 10.1109/TAI.2024.3515937.

R. Ronge is with the Munich School of Philosophy, Kaulbachstraße 31/33, 80539 Munich, Germany (e-mail: raphael.ronge@hfph.de).

B. Bauer is with the University Augsburg, Universitätsstraße 2, 86159 Augsburg, Germany (e-mail: bernhard.bauer@informatik.uni-augsburg.de).

B. Rathgeber is with the Munich School of Philosophy, Kaulbachstraße 31/33, 80539 Munich, Germany (e-mail: benjamin.rathgeber@hfph.de).

I. INTRODUCTION

The current landscape of Explainable Artificial Intelligence (XAI) is vast and can be confusing. There are a variety of XAI approaches [1]–[11] that are diverse and difficult to compare. Although the basic goal - to explain artificial intelligence - is common to all approaches, the means to achieve this goal are not homogeneous. For this reason, there are also a variety of taxonomies [12]–[16] that list these approaches and try to sort them into categories in order to bring some order to the field of Explainable Artificial Intelligence. All in all, much is left to be desired, as researchers find it difficult to sort through and search existing XAI solutions to find suitable candidates.

The aim of our paper is to provide a broader view of existing approaches, leaving the technical, functional domain and including a historical reconstruction as well as an analysis of the state-of-the-art using principles that we define (see figure 1). Through this broader analysis we aim to provide a better understanding of the prospects and challenges of XAI.

Starting with a historical reconstruction of XAI (see chapter II), we are able to pinpoint the source of the diversity of XAI approaches. We show that Explainable Artificial Intelligence is as old as AI research itself. While the goals of explanation were not originally synonymous with those of XAI today, it is important to understand XAI's close connection to AI even at its birth. In the face of ever-changing AI algorithms, explanations had to be adapted, leading to a variety of XAI approaches, which in turn instantiated different principles (see figure 1), resulting in the vast landscape of today. Given its chequered past, it is impossible to give a full account of XAI's history in this paper. Instead, we will outline the main lines of development and illustrate them with the most important examples. With this historical knowledge, we are able to develop a broad taxonomy that includes a wide range of dimensions (see chapter III-A), some of which have been proposed by others, while others are new. We illustrate our dimensions with meaningful XAI examples. An important difference between our taxonomy and many others are the non-technical dimensions (goal and target group), which again deepen the understanding of the XAI landscape and enable analysis of XAI beyond implementation differences. We do not want to limit XAI research to a small number of dimensions, but to show the variety and help to illustrate the possibilities of XAI. The shortcomings, on the other hand, are outlined in the chapter on paths through our taxonomy (see chapter III-C). In line with the rest of our paper, we do not attempt to be exhaustive - an endeavour doomed to failure in this rapidly changing field of research. Instead, we select the

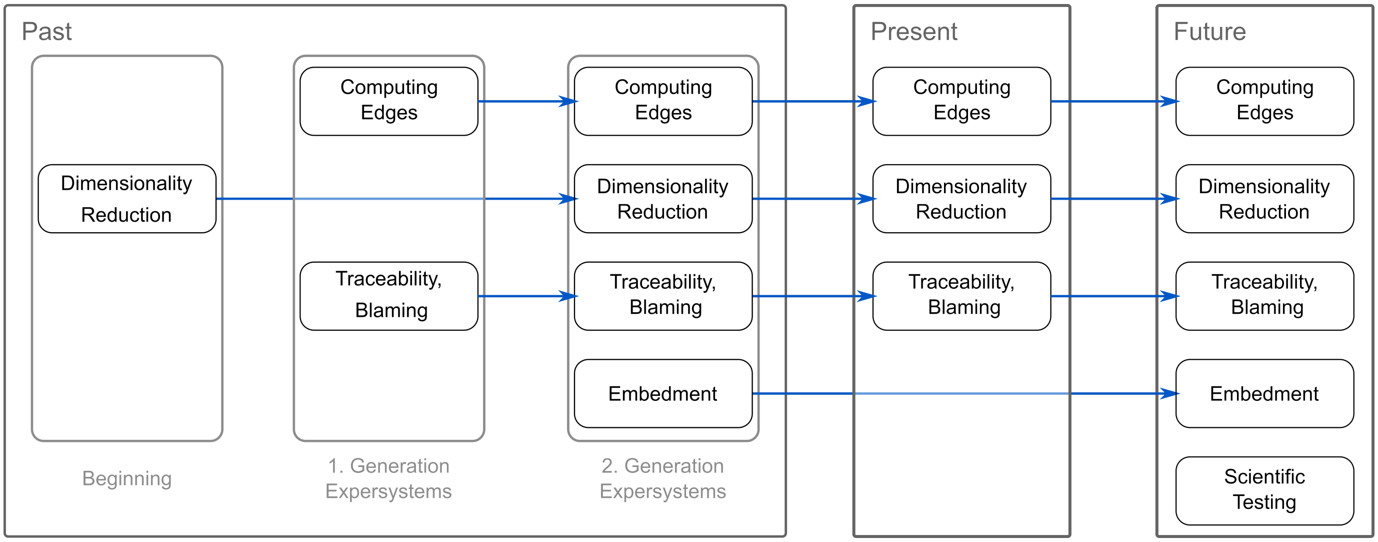

Fig. 1. Principles of XAI throughout history.

most telling examples to illustrate our taxonomy, as well as the most prominent shortcomings of current XAI research. The combination of historical reconstruction and our detailed analysis of our own taxonomy allows us to envisage principles that contemporary XAI already follows (see chapter IV) and those that might be helpful for future research (see chapter V). We do not claim our principles to be complete, but rather to provide a basis for discussion of future developments, which we mention in our conclusion (see chapter VI).

II. History of XAI

While the term 'Explainable Artificial Intelligence' (XAI) is relatively new, the concept is not. As long as Artificial Intelligence (AI) has existed, researchers have been interested in explaining the inner workings of AI and the concepts captured by AI models. The history of XAI is closely linked to the history of AI itself. Each new XAI era has been ushered in by a change in AI research. We are not able to give an exhaustive recollection of the history of XAI approaches. We refer the interested reader to the paper of Mueller et al. for a more detailed overview [17]. Our selected key XAI development strands show that XAI is by no means a phenomenon of the 21st century. The following historical reconstruction is aimed at the chapters on XAI principles (sections IV, V). There, we discuss principles of XAI history and state-of-the-art (see figure 1), as well as their relevance today.

A. *The Beginning (1940s – early 1970s)*

The birth of AI is inextricably linked to the birth of Explainable AI, long before that term was coined. However, the angle of attack of early explanations was very different from today. Frank Rosenblatt, the inventor of the perceptron, writes: "A perceptron is first and foremost a brain model, not an invention for pattern recognition" [18, p. viii]. This quote shows that explainability was of interest to early AI researchers in a very different way from today: "[The perceptron] is by no means a 'complete' model, [...] but it is, at least, an analysable model" [18, p. viii]. There was no problem with a black-box nature of AI models: AI was specifically designed to be a model of the brain and an explanation for brain processes that were otherwise too complex to understand. Regardless of the specific technical functionality, AI did not need additional explanations because it was built as an explanation of the human brain and its inner workings. This changed with the advent of Expert Systems.

B. *Expert Systems Explanations (late 1970s – mid 1990s)*

The symbiosis between explanations and AI models changed with the first AI winter, which ended in the late 1970s. AI research could not live up to its lofty claims, which led to the demise of research and ten quiet years [19, cf. e.g.]. After this period, AI research found its new application in practical models for the industry: Expert Systems (ES). These systems did not attempt to model the human brain, but relied on hand-crafted rules and used knowledge bases to solve specific tasks in their domain. With this new type of AI model, the need for explanations soon arose. Their development can be divided into the following two generations [17].

1) First Generation Expert Systems Explanations: The first Expert Systems were mainly used as (e.g., medical) advisors. One of the first "successful demonstration of scientific capability" [20, p. 19] of ES is DENDRAL, developed since 1965 at Stanford University [20]. However, it quickly became clear, that systems without a minimum of insight would not be of much use, as users would reject their results [17, pp. 44-45]. In this way, Edward Shortliffe's MYCIN became one of the most influential ES of the first generation [21]. Its "task [was] to assist with the decisions involved in the selection of appropriate therapy for patients with infections" [21, p. xiii]. MYCIN features an explanation component that is able to answer questions about its own knowledge base and specific consultation outcomes. Developing this question-answering process is a fairly straightforward task, as it involves only translating code rules (e.g., LISP commands) into human-readable text. There is an important caveat, however, which means that these "[systems,] which were developed to fulfil the need of end-users, may [often] have ended up being of greater value to developers" [17, p. 46]. Moore and Swartout called this caveat: "recap as explanation myth" [22, p. 11]. What they mean is that these systems "could not justify inference" [17, p. 47]. They simply recapitulated the rules used, without considering how humans (experts) might infer results, what knowledge they might already have and what rules they would normally use. This realization led to the emergence of the second generation of ES explanations [17, p. 49].

2) Second Generation: The Tutoring Approach: The flawed explanations of the first generation led the field of explanation in the very different direction [17, pp. 49-52] of so-called intelligent tutoring systems [23, p. 58]. Although they overall were solving problems that are not comparable to today's XAI challenges, some very interesting ideas were implemented that improved on first generation ES explanations. Tutoring systems were no longer stand-alone Expert Systems, e.g., for medical diagnosis, with an optional explanation component. They were intended to help train doctors to improve their diagnostic skills [17, pp. 49-52]. This new focus led to the creation of user-models. Second generation system explanations "accounted for the student's recent behaviors and claims" [17, p. 52] and "often focused on making explanations context-sensitive" [17, p. 52]. William Clancey implemented these ideas in 1982 with GUIDON, an ES that uses the knowledge base of MYCIN and has an additional 200 separate tutoring rules [24, p. 8]. These tutoring rules implement a student model that keeps track of the MYCIN rules that a student knows, is likely to be able to use for a given case, or has already applied [24, pp. 10-11]. Having a student model is of paramount importance, as "simply representing in an ideal way what to teach the student is not a trivial, solved problem" [24, p. 14].

In addition to the focus on user interaction, the second generation of explanations explored the rules themselves. XPLAIN by Swartout combats the "recap as explanation myth" [22, p. 11], with meta-rule-learning [25]. It works by defining a domain model (i.e., textbook facts) and domain principles (i.e. domain heuristics/rules) and then using an automatic programmer to create the rules of the ES, thus making them context-sensitive [25, p. 287].

A third important concept, developed for the second generation is counterfactual explanations [26, cf. e.g.]. They were found to be "features needed in generating an acceptable explanation" [27, p. 19] and were therefore used in BLAH from 1980 by J. L. Weiner. Even today, authors argue – similarly to Weiner – that when a user asks 'Why?' what is often meant is 'Why X instead of Y?' [28, p. 16], which makes counterfactuals a good option for answering these questions.

Of course, these concepts are not the only ones in the second generation of ES explanations, but they are the most promising ones to be transported into today's XAI research (see section V).

C. Rebirth of Explanations with Neural Networks (early 2010s – today)

The second AI winter with its collapse of the ES market was followed by a much longer "Explainability Winter" [17, p. 60] (late 1990s – early 2010s) with little to no research into new explanation methods for a decade. It was only after the comeback of AI in the form of Neural Networks (NNs) in the early 2010s that explainability research was reinvigorated. Newly adapted training algorithms and ever-increasing computing power made this new era of AI performance possible. NNs became the workhorses of today's AI research, but since the rebirth a wide variety of AI solutions have evolved.

These new developments, however, had transparency drawbacks that had to be addressed by new research into explainability. While the first AI approaches were designed to be understandable models of our complex brains (section II-A), and ES worked with humanly understandable concepts or rules (section II-B), today's ML models achieve the highest performances because of their massively parallelized structures. As they are self-learning, human insight is already minimal. And with larger models, the insight only gets smaller and smaller, and today, even developers are unable to achieve previous levels of understanding. The current XAI problem is no longer one of translating and representing rules, but of finding and extracting them. These difficulties have led to a new strengthening of explainability research, the coining of the term XAI since the 2010s, and the crystallization of XAI as a new research area [17, pp. 62-66].

III. State of Explainability Research

Compared to earlier phases of XAI, today's challenges are fundamentally different. In the past, the training of an AI model was minimal – if present at all. While the rules of an ES were created by humans and 'only' needed to be translated into English, NNs, for example, only implicitly retain these rules. Finding and extracting rules from AI models is not an easy task. That is why, until recently, XAI has focused mainly on the models themselves and less on human interaction with explanations.

Based on the historical reconstruction in the chapter above, we can use the analysis of the state-of-the-art in the following to get a clear understanding of XAI's prospects and challenges.

Today, XAI has an immensely diverse ecosystem. This is due to two main factors. First, it is based on a wide variety of AI models with many different data types and multiple learning algorithms. In earlier AI models, the input data was textual or numerical. Today, AI models can also handle visual and auditory data. This presents new challenges, but also new opportunities for presenting explanations to users. Second, the diversification is encouraged by the current development method: new explanations are developed on a case-by-case basis, e.g., for a specific model, use-case, or user and often don't take into account larger methodological considerations.

The following sections analyse the XAI landscape in two ways. First, approaches are categorized along different dimensions of goals, technical applications and user interaction (section III-A). By systematizing the landscape, it becomes possible to discern approaches and to identify their commonalities. Second, four XAI examples and their taxonomy paths are given to hone the taxonomy and to highlight some key shortcomings of today's XAI (section III-C). The results lead to the definition of principles in the next section (section IV), to help find common features as well as gaps in XAI, that might be important for future research.

A. Taxonomy

The taxonomy below (figure 2) features a kind of flow pattern to illustrate the different steps taken by an XAI method or during its development. This leads to a hierarchy of dimensions: from start to finish, the dimensions are sorted from foundations to computational methods to output modalities. Each XAI instantiates a path through the dimensions and categories. Since categories are not always exclusive, an XAI can incorporate multiple.

The myriad of XAI approaches leads to a variety of different ways to taxonomize them. After presenting our taxonomy in the following, we will give an overview of other taxonomies and compare them to our approach.

While *goal* and *target group* are preliminary dimensions and can be decided outside of a specific technical implementation, *integration stage*, *applicability*, *methodology*, and *scope* concern the nuts and bolts of an XAI approach. The last three dimensions - *result modality*, *result format*, and *presentation format* - categorize the results of XAI approaches.

1) Goal: The most fundamental dimension of an XAI approach is its *goal*. Every other dimension depends on it in one way or another. XAI goals are the translation of either intrinsic desires or external constraints. Intrinsic desires can be human curiosity, societal acceptance of an AI model or knowledge extraction, while legal compliance, security, and debugging of the models are external constraints that require the use of XAI. The goals of an XAI given by authors of XAI approaches are regularly very broad, mentioned in passing and not followed up on. This is due, at least in part, to the difficulty of proving goal fulfilment, which can be a very domain- and application-specific endeavour. A second important reason why stated goals often remain vague, is a mismatch between the supposed embedding of a goal and the actual results of an XAI. XAI papers typically do not define the term 'explanation' [9], [29, cf. e.g.]. When they do, it is a working definition for themselves, which lacks theoretical grounding [12]. Given the goals that many researchers name, their working definition of 'explanation' seems to rely solely on human interaction. The only way to prove goal fulfilment would be empirical studies showing that an XAI can fulfil a claimed goal for a given user. Instead, the researchers rely on our imagination to perceive goal fulfilment simply by understanding the mechanics behind the XAI.

2) Target Group: An XAI approach must meet different requirements, depending on the *target group* it is aimed at. Target groups can be distinguished by their domain knowledge of AI and its application domain, as well as their use of explanations. The following target groups are an adapted version of the human-centred taxonomy by Langer et al. [30]. While the *subject* (someone who is affected by the output of an AI) typically has neither ML knowledge nor domain knowledge, a *developer* has a large knowledge base in ML, just as a *deployer* knows the ins and outs of his domain. Given these differences in prior knowledge, explanations need to provide different levels of detail. Especially because these explanations will be used either to improve an AI model (*developer*), to get insight-driven explanations for AI results (*user*), to assess performance (*deployer*), or to get a knowledge base on which to decide if an AI model is fair (*regulator*, *subject*). These five target groups are not set in stone, and in some application cases, multiple groups may be embodied by a single person.

3) Integration Stage: The first technical dimension to group XAI approaches by is *integration stage*. Either an AI model is developed in such a way that explanations can be obtained directly from it (*ante-hoc*), or an already existing opaque model needs some kind of addition making it explainable (*post-hoc*). For *post-hoc* methods, the performance and interpretability of an AI model is a more flexible two-step process. Typically, models in the first group are so-called transparent models (e.g., decision trees (DT)), for which the computational process is relatively simple and semantically meaningful. While there are applications where it is useful to rely on this method instead of building more complex models [31, p. 5], it is by no means a solution for all cases. A transparent model is not necessarily understandable for two reasons. To be understandable the input features must be meaningful to humans, and the model must not be too large, otherwise the AI model's decision process cannot be followed by a human [32], even if it is semantically meaningful. *Ante-hoc* methods exist only with a specific AI model, as it is built in a way that requires no further explanation. Therefore, a categorization by *applicability* and *methodology* is only meaningful for *post-hoc* models.

4) Applicability: Post-hoc explanations are the more common solution because they can be used on already existing AI models. They can either have a *broad range* like SCOUTER [33] because they are data-oriented, or they are more dependent on the AI model structure itself and therefore have a *narrow range* like DeConvNet, which is only applicable to Convolutional Neural Networks (CNNs) [34]. The two extremes of this dimension are *model-agnostic* and *model-specific*, but most approaches lie on a spectrum in between. *Post-hoc* approaches that are completely *model-agnostic*, are, e.g., LIME [9] and SHAP [35], which only alter input data to measure a difference in output. However, the connotation of model (in)dependence is multifaceted. Does '(in)dependence' mean that an XAI is applicable to every AI or to every AI of a specific type? Is it usable only with knowledge about or access to the inner computations, or without it? Does it require an alteration of the structure of the AI model or not? This shows that the groups in this dimension are not as clear-cut as in others, and the categorization is more of a broad direction.

5) Methodology: The dimension of *methodology* is interconnected with the *applicability* dimension above. Different methods have a *narrow* or *broad* application range based on their functionality out of four different methods. First, there are XAI approaches that modify [36] or add to the

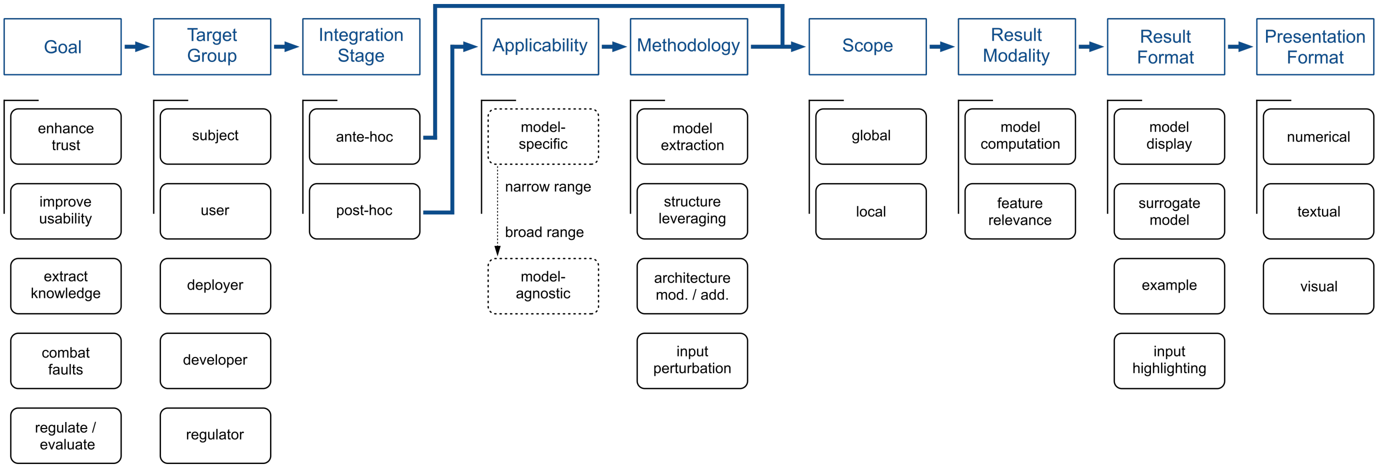

Fig. 2. Taxonomy: XAI State-of-the-Art

structure of an AI model [37] (*architecture modification / addition*). Second, *structure leveraging* methods often use the gradient information of a NN to determine the importance of an input feature for the AI model's output [38]. Some methods exploit the structure of AI models by visualizing data at network nodes, e.g., DeepVix [2]. Third, *input perturbation* approaches are typically found in *model-agnostic* XAI. Input data is altered, and the output is observed to gain information about the input-output-relation of the AI model [9]. Fourth, XAI methods can work through *model extraction*. A second – interpretable – model is trained to mimic the results of the first opaque model, with the expectation that it learns the same decision boundaries [39].

6) Scope: The dimension of *Scope* is frequently utilized in XAI taxonomies [40]. At this point it is also meaningful to reintroduce *ante-hoc* methods. The two categories describe whether an XAI approach explains individual instances (*local*) or the entire model (*global*). *Post-hoc* approaches usually fall exclusively in one of the two. However, some *local* methods can be expanded to approximate *global* ones [9]. A transparent *ante-hoc* model always exposes its structure. The size of which determines whether it can be understood as a whole or only in terms of instances' paths.

7) Result Modality: As demonstrated previously, XAI can be classified based on their *goals* and technical configuration. Additionally, various types of it results and result presentations can be distinguished.

There are two *result modalities*: an XAI can either explain through *feature relevance* or by providing insights into the AI *model's computation*. *Feature relevance* explanations aim to highlight the significant parts of the input for a given output [29]. *Model computation* approaches aim to explain a model's function by focusing on its inner computations, without referencing a single instance. For example, Springenberg et al. visualize different layers in a Convolutional Neural Network (CNN) to achieve this [41].

8) Result Format: XAI with both modalities can present results in different *result formats*. It can explain either by using *examples* drawn from the training data set [42, cf. e.g.], by visually *highlighting* parts of input data (images, tabular data, or text) [38, cf. e.g.], by displaying (parts of) the *model* [43, cf. e.g.], or by creating a *surrogate model* that is transparent [44, cf. e.g.]. *Feature relevance* approaches typically use *highlighting*, while *model display* or even a full *surrogate model* are more effective to explain an AI *model's computation*.

9) Presentation Format: The final dimension by which XAI approaches can be grouped is their *presentation format*. The choices of format heavily dependents on the earlier dimensions, as well as the AI model and data type. For instance, *visual* explanations are typical for XAI approaches that try to *locally* explain the *feature relevance* by *highlighting* the input of an image classification algorithm [29], [33], [45, cf. e.g.]. However, some use *visual* presentations to provide explanations by displaying *surrogate models* globally [46, cf. e.g.]. *Textual* explanations are used to explain the 'reasoning' of AI in sentences to make them as easily understandable as possible [47]. The *numerical presentation format* is commonly used for XAI approaches aimed at *developers*. This format can provide a more nuanced result, but may only be understandable by someone with prior knowledge [48].

B. Literature Review

In the following, we present some other taxonomies that are relevant to our approach. This will give an overview of the state-of-the-art and further motivate our own approach.

As mentioned above, most of the dimensions are not new, but some are more widely used than others. For example, Lipton in 2018 [32] and Tomsett et al. in 2018 [49], while not being explicit taxonomies, mention different dimensions in their descriptions of the research field. Lipton first analyses the notion of transparency, distinguishing between simulatability, decomposability and algorithmic transparency [32, p. 40]. Basically, we are splitting our category of *ante-hoc* approaches. Lipton distinguishes another type, *post-hoc* approaches, which he further subdivides into *textual* and *visual* explanations and explanations by *examples*. [32, pp. 40-42]. Both categories are regularly mentioned by researchers. Tomsett's paper, on the other hand, is one of the few examples that analyses our dimension of *target group*. They identify six target groups:

creator (organisation that owns the model as well as the developer, in our approach *developer* and *deployer*), operator (person who works with the model and provides its inputs, in our approach *user*), executor (person who acts on the outputs of the model), in our approach also *user*), decision-subject (person affected by the output of the model, in our approach also *subject*), data-subject (person whose personal data was used to train a model, in our approach also *subject*), examiner (e. g., auditors, in our approach *regulator*) [49, pp. 9-10].

One taxonomy that names a variety of categories that are present in other taxonomies and are also included in our approach is that of Arrieta et al. from 2020 [50]. They mention the distinction between *model-specific* and *model-agnostic* approaches. They further distinguish *model-agnostic* explanation by simplification (which partly overlaps with *model extraction* of our taxonomy), feature relevance explanation, local explanation and visual explanation [50, pp. 92-94]. While these are all valid categories of *model-agnostic* approaches, in our taxonomy they fall into different dimensions. Our flow-like taxonomy helps us to distinguish between the different qualities of the categories and saves us from juxtaposing them. This becomes even clearer when looking at the tree-structured overview of Arrieta et al. [50, p. 93]. One can see the mentioned categories at different levels of the tree for both *model-agnostic* and *model-specific* approaches. While this is a correct assessment, it is not particularly good for clarity.

A more technically focused taxonomy can be found in the paper by Guidotti et al. [51]. They review the state-of-the-art and divide it into four categories: model explanation, outcome explanation, model inspection, and transparent box design. Model explanations are what we call *global* in *scope* and stick to *model extraction* as their *methodology*. Outcome explanations are simply all *local* explanations. Model inspection describes *post-hoc* approaches and transparent box design *ante-hoc* approaches respectively [51]. Again, these four categories have different levels of detail (and belong to different dimensions in our taxonomy). Additionally, they give examples of XAI for each category and define the features that these examples could have. Related to our taxonomy, they propose the feature of generality, which corresponds to our dimension of *applicability*, and randomness, which we describe by the category of *input perturbation*. Again, discussing these in a comparative way seems somewhat confusing. However, Guidotti et al. supplement these features with others that are explicitly aimed at computer scientists. They mention the exact ML-model as well as the data modality. And they go on to say whether an approach gives specific examples and whether it has an accessible code base and dataset. This could help computer scientists to choose an XAI approach that suits their purpose.

A taxonomy with a specific research area, namely clinical applications, as its background is Antoniadi et al. in 2021 [52]. They review a variety of XAI approaches in medicine and discuss them in the context of clinical decision support systems. In the end, they sort them into three dimensions that overlap with ours: model-agnostic/specific (our dimension of *applicability*), ante-hoc/post-hoc (*integration stage*), local/global (*scope*) [52, p.13]. Although they do not name their dimensions in the same way, they are very similar to those proposed in this paper. In addition to dividing the field into many more dimensions, we analyse the dimension of *applicability* in more detail, as we find a binary distinction too restrictive.

Timo Speith in 2022 [40] and Gesina Schwalbe and Bettina Finzel in 2023 [15] attempted to unify the taxonomy landscape. While Schwalbe and Finzel devise a taxonomy that incorporates different approaches, Speith extends a taxonomy with an XAI database and the idea of a guiding decision tree. Speith focuses on the technical application of his approach. It is designed to assist researchers in selecting an XAI approach. Schwalbe and Finzel, on the other hand, attempt a meta-categorisation. At the highest level, they distinguish three aspects of XAI methods: problem definition, explanator, and metrics [15]. While they mention their chronological order, they later present them collocated. This in turn obscures the order of the subcategories of all three aspects. The resulting taxonomy by Speith is similarly extensive, but mainly mentions the technical aspects of XAI approaches [40, p. 2246]. Consequently, the clusters mentioned by Speith overlap significantly with ours (but lack the flow-like structure): stage (in our approach *integration stage*), scope, functioning (*methodology*), result (*result format*), output format (*presentation format*) [40]. For Schwalbe and Finzel, the subcategories of the explanator aspect in particular overlap with our dimensions [15]. As both taxonomy surveys focus on the technical functioning of XAI approaches, our first two dimensions *goal* and *target group* are missing in both resulting taxonomies.

To illustrate the advantages of our approach, we will now explain our taxonomy, its dimensions and categories, and its flow-like structure by incorporating some XAI examples into our taxonomy.

C. Taxonomy Paths

The previous section provided an overview of all XAI dimensions and their categories. The taxonomy provides an effective way to conceptualize the current XAI landscape, although new developments may require adaptation. However, the wide variety of XAI approaches makes it difficult to develop a precise understanding of the categories and their meanings for each approach. To improve this, and to enable the identification of current overarching issues, missing principles, and potential solutions, the following three sections illustrate the structure and meaning of the taxonomy by tracing the paths of four selected XAI approaches through it. The different paths highlight not only differences but also commonalities of approaches similar to the archetypal examples selected below.

1) 'Why Should I Trust You?' by Ribeiro et al.: One of the best known XAI algorithms is LIME (Local Interpretable Model-agnostic Explanations) by Ribeiro et al. [9]. The path through the taxonomy is shown in the figure 3. LIME creates new (pseudo) instances by perturbing the input in the region of the original instance. By observing the changes in the output, a linear approximation is computed with the created instances weighted by their closeness to the original instance. In this way, it is completely independent of the model itself and *model-agnostic*. The coefficients of this linear approximation are then used as an input importance measure for the output at the given instance [9, pp. 3-4]. Ribeiro's paper is one of the few that also scientifically address the first two dimensions of *goal* and *target group*. Several user studies are conducted to verify that users are able to *extract knowledge* and *combat faults* (improve the model) based on LIME [9, pp. 7-9]. However, they prove their main goal to *enhance trust*, only by conducting a simulated user study (with positive results). The possible *target groups* of LIME are not clear. Based on the wording and studies conducted, one can infer that LIME is aimed at *deployers*, *developers*, or *users*. On the technical side, since LIME can be applied to any existing model, it is a *post-hoc* approach. The technical description of LIME above shows that it is *model agnostic*, which is achieved through its strict use of *input perturbations*. The normal mode of operation of LIME is to produce *local* explanations (see [9, pp. 5-6] for a possible extension). The results then display the importance of individual input features (*feature relevance*, *input highlighting*. This is done either by highlighting so-called super-pixels in images (*visual*) or by presenting a bar chart of the importance (e.g., of words) to the user (*numerical*).

2) 'XAI for Transformers' by Ali et al.: A relatively recent XAI approach comes from Ali et al. in 2022 [38]. They adapt the LRP method [45] for transformers. They found that transformers, unlike, e.g., feed-forward NNs, cannot satisfy the propagation constraints required by known relevance approaches. The Ali et al. path shows a different type of XAI (see figure 4). It also presents some problems that many XAI papers have. First, *goal* and *target group* are not as clear as before and much less empirically verified, as they focus only on technical aspects of their approach. They claim a "fairly intuitive explanation method" [38, p. 1] and based on the nature of their results, it is inferable that the method could be either used by *regulators*, *developers*, or *users*. The suitability for *regulators* also stems from the fact that Ali et al. want to be able to "verify whether the model makes fair decisions and does not discriminate protected classes" [38, p. 1]. Therefore, it seems to be reasonable that *regulation* and *evaluation* is their *goal*. Their *post-hoc* explanation method has *narrow range*. It is possible to use it on more than one specific model, but it is specifically designed for the transformer model class. The method employed is *structure leveraging*. It is an extension of LRP, which itself uses the structure of a Neural Network to assign higher layer relevance to lower layer neurons. This is done to obtain a *feature relevance* measure for a single input image (*local*). The relevance is then displayed as a heat map over the original input image to highlight the important areas (*input highlighting*, *visual*).

3) 'Generating Visual Explanations' by Hendricks et al.: Another typical example of XAI research papers following a different path is the explanation by Hendricks et al. [53] (see figure 5 for path). They start with an opaque CNN and add a Long-Short-Term-Memory (LSTM) extension [53, pp. 3-4]. They design a new loss to retrain this combination correctly. After retraining, the CNN is used to process an image and predict its class, while the LSTM produces a natural language explanation of the image features that led to the classification at hand [53, pp. 7-9]. Again, there is little information about *goal* and *target group*. On the one hand, they mention that they want to be able to "understand network mistakes and provide feedback to improve classifers" [53, p. 3]. However, they claim that their approach is "useful for non-experts" [53, p. 4]. In addition, "understanding and interacting with AI systems" [53, p. 3] should be encouraged by their explanation. All in all, it might be possible to claim that *knowledge extraction* and *combating faults* are the goals of Hendricks et al. and their *target groups* are either *users*, *subjects*, or *developers*. None of their claims are supported by any kind of user study. Hendricks et al. only check their generated sentences theoretically [53, p. 11-16]. As Hendricks et al. start with an opaque CNN it is a *post-hoc* approach. The specific structure of the CNN is irrelevant, which makes it a *narrow range* approach. The methodology they follow to achieve their explanations is, as explained above, an *architecture addition* (adding an LSTM to an existing CNN). The LSTM than produces *local feature relevance* measures that highlight parts of the input (*input highlighting*) as *textual* output.

4) 'Interpreting Blackbox Models via Model Extraction' by Bastani et al.: The model extraction algorithm by Bastani et al. [46] is a good example to showcase a very different path through the taxonomy (see figure 6) that still has some of the problems as the two examples above. Their proposed algorithm extracts a (greedy) decision trees from an existing model by sampling input data based on the training set distribution. First, they explicitly state that their focus is to "enable data scientists [(*developers*)] familiar with machine learning to understand and validate the complex model [(*extract knowledge*)]" [46, p. 5]. Second, as the tree building algorithm is used on the outputs of existing models and only uses the input-output relation, it is, like LIME, a *post-hoc*, *model-agnostic* XAI. As Bastani et al. state in the title, they propose a *model extraction* algorithm. Building a transparent model that mimics the behaviour of the original tries to explain *global model computation*. Finally, while they could present their results in multiple ways, they select to display the extracted model (a *surrogate model*) either whole *visually* or its corresponding rules *textually*. However, their claims are only partially substantiated. While they show that people are able to interpret the resulting decision trees, they do not show how this helps to understand and validate the underlying models. Furthermore, while the decision tree approximates the best greedy decision tree, this is no guarantee that the underlying model works in the same way.

IV. Principles

The last two sections have highlighted the chequered history of Explainable AI and the current diversity of approaches. The following section now analyses the commonalities and gives them a firm grounding in three principles.

The notion of principles is very helpful at this point. As mentioned earlier, XAI research is diverse not only because of a multitude of AI models but also because of its development procedures, which often have no common ground. This leads to a field of research whose taxonomy has multiple dimensions and many more subcategories. Principles help us to abstract

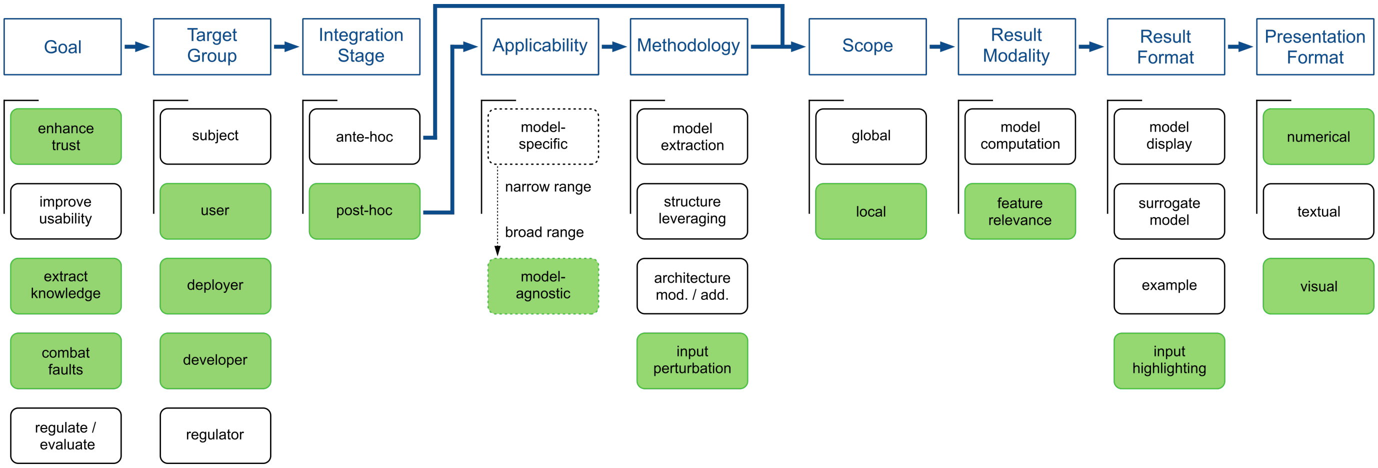

Fig. 3. Taxonomy Path: LIME

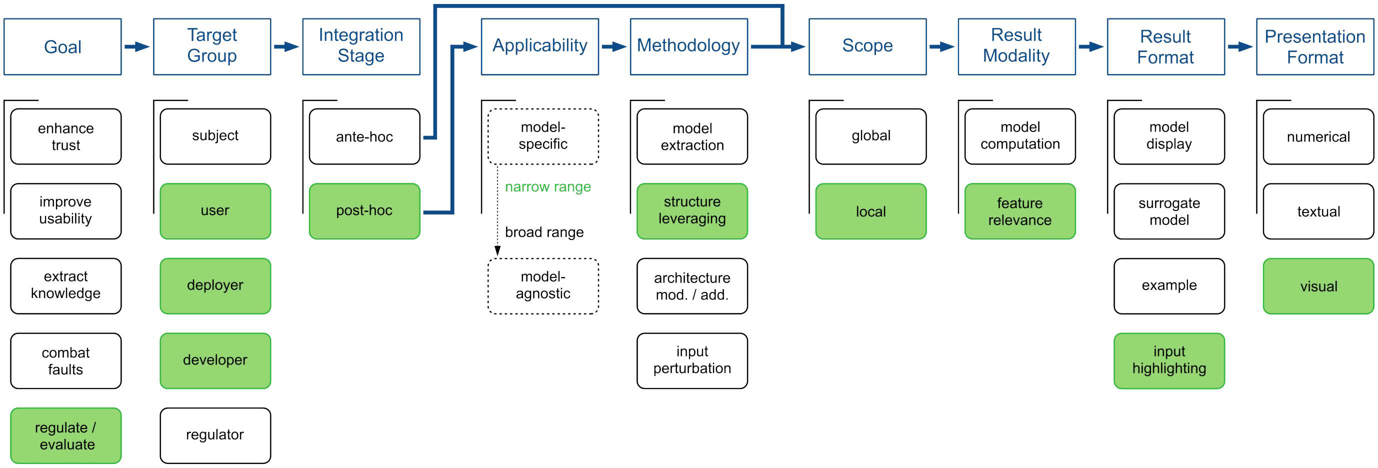

Fig. 4. Taxonomy Path: XAI for Transformers

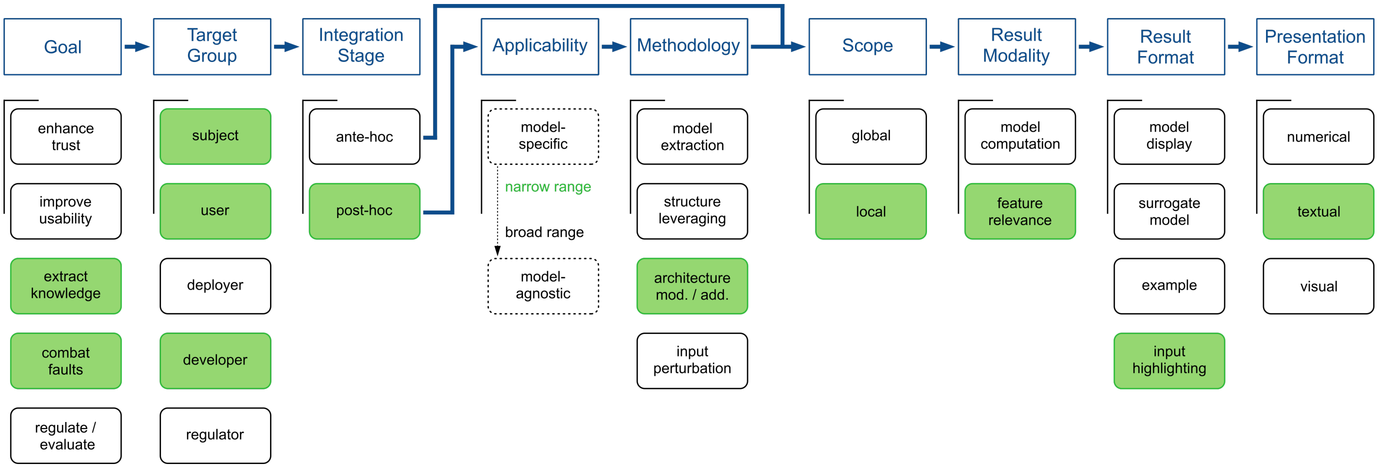

Fig. 5. Taxonomy Path: Visual Explanations

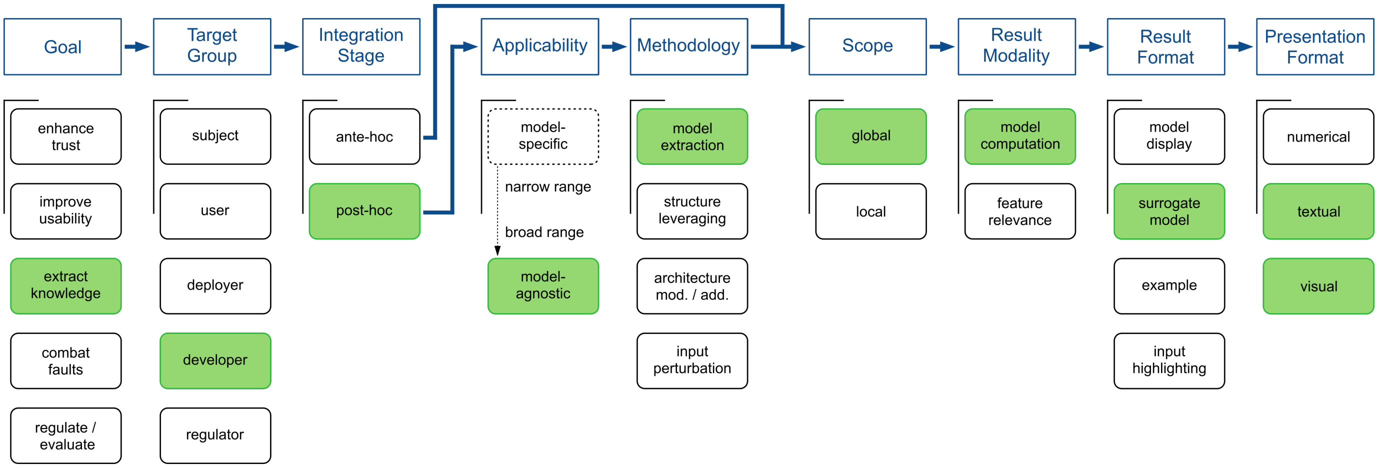

Fig. 6. Taxonomy Path: Model Extraction

from the purely technical level and to find common concepts between approaches across dimensions. Moreover, principles create a system at a higher level allows analysis of the embedment of approaches in different notions of explainability.

The principles below represent some of the defining strands of XAI today. However, linking the principles back to earlier research will help to find principles that are underrepresented in or missing in contemporary XAI. The principles are neither exclusive nor a complete picture. They are intended to stimulate discussion and provide an important perspective on the development of XAI. Similar to Buchholz (2023) [54], our analysis is two-sided. On the one hand, our taxonomy chapter in particular is descriptive. The following principles have the same function and describe the field of XAI research. On the other hand, we argue why XAI necessarily conforms to these principles and should conform to the principles proposed in the next chapter (chapter V). We show how all the different paths through the taxonomy necessarily adopt one or more of the following principles, and why our proposed principles would improve XAI research.

A. Computing Edges

Most XAI approaches try to compute some kind of edge. This is partly due to the fact that most XAI is intended for classification models. The idea is to use the distinction between decision regions to either explicitly or implicitly explain a result compared to a possible alternative. This fits quite well with a finding by Tim Miller. He analysed explanations from a social science perspective and found that AI explanations should be contrastive [28, p. 6].

1) Presence in History: Throughout the history of XAI there have been approaches that have followed this principle. As Expert Systems use decision rules, explanations in the first and second generation consisted, at least in part, of translating these rules into text, to explain the decision boundary embodied in the rules. A novelty of current XAI, however, is that these rules must first be approximated.

2) Decision Boundary: All *ante-hoc* (see *integration stage* in section III-A3) and *model extraction* (see *methodology* in section III-A5) explanations basically do some kind of decision boundary detection, as a transparent model (e.g., a decision tree) shows the switch points of a model. The best example would be a linear model, whose visualization is typically just the decision boundary. Other XAI methods do not necessarily display the decision boundary, but still use it to compute their explanation. For example, *input perturbation* methods implicitly rely on a crossing of the decision boundary, as they use the information about the change in the prediction due to a perturbation (see, e.g., LIME [9]). Independent of *integration stage*, *applicability*, and *methodology*, XAI based on counterfactuals depends differently on the decision boundary. A counterfactual is a (synthetic) instance with changed values for some features to show what (and how) features of an original instance need to be changed to produce a different result. This is basically moving an instance to the other side of a decision boundary.

B. Dimensionality Reduction

A second defining principle is dimensionality reduction, or simplification. Today's ML models are so complex, that it is necessary to reduce their complexity. This principle is directly at work in *post-hoc* explanations while *ante-hoc* models have to follow it during the design process. The different ways of achieving this reduction are explained below.

1) Presence in History: This principle was already present in the very first AI models. However, as explained above (Chapter II-A), explanations worked the other way around. The first AI models were explanations of the brain itself, and they tried to do this by simplifying the way the brain worked immensely. Later, the failure of the first generation of ES explanations can be partly attributed to their failure to follow the principle of dimensionality reduction. Simply recapitulation decision rules could lead to overly long explanations paths, cluttered with implementation details and of little help to the user. [22, p. 12]

2) Scope Reduction: Each XAI approach must perform at least one type of reduction: If the *global model computation* is explained, necessarily the influence of single instances is omitted and vice versa for *local* explanations. However, there are drawbacks to be aware of. For example, a global model approximation can lead to low fidelity and a local approximation is only valid for its neighbourhood.

3) Model Reduction: The reduction of dimensionality for model computation is already somewhat present in the computing edges principle. Focusing on the edge of a class is a simplification because it omits computational differences between instances of a class. Instead, it focuses on the important classification differences between classes and is a simplification of the input-output relation. With *model extraction* algorithms, a complex model is simplified by replacing it with a transparent one with nearly identical behaviour. The new model operates with fewer steps compared to the layers in a Neural Network, for instance. It also reduces the computational load (less parallelization of a Neural Network or a random forest). This goes hand in hand with a better traceability (see the next principle IV-C).

4) Input Reduction: In addition to reducing in model computation, it is also possible to reduce the input or feature space. *Feature relevance* approaches add a lot of information about the importance of different features. However, in the end they divide input data into positive and negative influence groups. This is an example of simplification, as it does not require the user to deal with a large number of input features and allows them to focus on the most influential ones.

An XAI approach that combines several reduction methods is counterfactuals. The first method is model reduction, as a counterfactual focuses on the decision boundary rather than the full range of a class. It also simplifies through its implicit *feature relevance*. Simplification occurs because the influence of most features is not shown to the user, as they only have access to a few modified ones. The final dimensionality reduction method that is implemented by counterfactuals is, by their very nature, *examples*. Examples use the human ability to recognise patterns as a basis for explanation. A few class examples that are representative of the whole class (reducing the input space) do not need an explicit proximity measure. Their explanation is based on the fact that humans can detect similarities and differences fairly easily and are able to identify the defining characteristics.

C. Traceability, Blaming

If the focus of an XAI approach is *model computation*, this last principle can be called traceability, otherwise blaming, if it is a *feature relevance* approach. In both cases the goal of the explanation is to attribute importance (blame) for a result to parts of the model or input. At their core, traceability approaches attempt to solve the explanatory relevance problem of well-known scientific explanation models [55, cf. e.g.]. Although this task is easier to solve in the confined space of a ML model, it is not a trivial task.

1) Presence in History: Traceability was the overriding principle that guided the explanations of the first generation of ES. Their main aim was to translate the rules by which an input was processed into human-readable language. They provided traceability by showing the steps taken.

2) Model Traceability: The difficulty of achieving full traceability increases with the size of the model. A transparent *ante-hoc* model can have traceability built in. It is simple and small enough to be able to follow a path through the entire model. However, an AI model can be so large that tracing one path would be meaningless. XAI methods, therefore, use snapshots at critical points in the network to gain understanding. This is true for all the methods that visualize Convolutional Neural Network filters [41, cf. e.g.]. *Structure leveraging* XAI bridges traceability and blaming. LRP [45] and its derivatives [38, cf. e.g.] trace paths through a network, but display their results only as heatmaps of the input image, which is the working mode of feature blaming.

3) Feature Blaming: The way *input perturbation* methods follow this principle is not by tracing a path from input to output, but by attributing blame to input features based on their influence in shifting the output. In this way, they identify the most important input-output relationships. This is also true for counterfactuals and their feature blaming mechanism for moving an instance to another class.

V. Future Principles

Explainable Artificial Intelligence has changed several times in the course of its development (see section II). Often new AI models were the trigger for this development. However, the change from first to second generation explanations for ES was mainly due to the shortcomings of the explanations. This points to the question, if a similar shift might be fruitful today? Current XAI research consists of a wide variety of approaches. The taxonomy in section III-A establishes categories to sort this vast landscape.

The flow-structured taxonomy, together with our reconstruction of the history of XAI, led us to our selection of proposed principles. The principles can help to analyse XAI research by embodying a higher-level structure. The fact that our principles are based not only on the current state-of-the-art, but also on our analysis of the past, gives them the necessary foundation in this fast-developing field of research. However, the different paths in section III-C show that current XAI approaches already follow a variety of paths. A taxonomy with some claim to generality cannot cover every possible edge case without becoming arbitrary. As a consequence, our principles may miss some (current or future) XAI approaches. In what follows, therefore, we propose two additional principles that we believe could help future developments of XAI. They are an attempt to bridge the gap between current principles, to capture more XAI approaches, and to stimulate discussion about future XAI developments.

A. Embedment

The main principle discovered to be necessary for the second generation of ES explanations was, best described as *Embedment*. The problems of the first generation were labeled by Moore and Swartout as the “recap as explanation myth” [22, p. 11]. It was combated in two ways: second generation explanations began to “[account] for the student’s recent behaviors and claims” [17, p. 52] and became “context-sensitive” [17, p. 52].

1) user interaction: Clancey's GUIDON specifically incorporated *Embedment* for user interaction and attempted to solve the problems of MYCIN [24]. Clancey used a system that kept track of a student's knowledge and allowed the system to tailor its explanations to the user and his prior knowledge. This shows the importance of defining a *target group* for an XAI approach. Target groups are differentiated by their desires and their prior knowledge of ML and application domains. Designing an XAI for a specific *target group* enables leveraging prior knowledge. The realization that explanations need to be more than a list of rules has led to the development of systems that incorporate knowledge that was previously implicit. Neches and Swartout use procedural and declarative knowledge about a domain in their 'Explainable Expert System' [56]. Their idea was to make available the knowledge that went into development. Today, this would mean using knowledge that was implicitly learned by the model during training. In practice, one would have to at least document the training procedure, its steps, and data. The interaction mechanisms developed for second generation explanations were later summarized by NUCES [57]. The user interaction mechanisms it applied consisted of a multimodal knowledge graph that was accessible to the user, incorporating *visual*, *textual*, and *numerical* information. Today, unlike the first generation ES, we are already able to produce *visual*, *textual*, and *numerical* explanations. There are even some frameworks that incorporate multiple XAI approaches [58], [59, cf. e.g.]. In the future, we need to deploy more multimodal explanations that structure different modalities in a meaningful way.

2) Context Sensitivity: In addition to user interaction, NUCES is also context sensitive. The reason why the second generation of ES explanations followed context sensitivity is due to the fact that they were tutoring systems and their explanations had to be linked by their very nature to the context of their models. Today there are decidedly model-agnostic XAI approaches [9, cf. e.g.], and even those with a more *narrow range* are intended to be applied to a variety of models [38, cf. e.g.]. Context sensitivity is not an option for these approaches. However, even for today's XAI there exist some research regarding the importance of context-sensitivity [60, cf. e.g.]. Stieler et al. developed a domain-specific approach that takes into account the important factors for skin cancer classification. The context sensitivity of NUCES also applies to the available data, as it is possible to explore the entire knowledge tree. For today's AI models, this is no longer possible due to the training done during development. Typically, one does not have access to that information during deployment. Without the knowledge about what the models had access to, a grounding to ground truth is not possible: "[How] can we evaluate the importance of someone's salary to a loan decision, if the classifier can only evaluate people with valid salaries" [61, p. 282]? A context sensitive explanation would make it easier to identify possible knowledge gaps in the model's reasoning.

B. Scientific Testing

Interestingly, there is one principle that is often missing from today's XAI, as well as, second generation ES explanations: *Scientific Testing*. This is a principle that is not directly related to the functioning or results of an XAI, but to its overall development. Section III-C mentions that many approaches are not scientifically tested. Even if they are, there is no generally accepted measure of 'explanation goodness'. In addition, the tests are only performed for the specific use case of the authors (case-by-case development) and not on a collective data set. The combination of these issues leads to undefined performance and, in the worst case, XAI approaches that do not work as they should. For example, Adebayo et al. showed in 2018 that some *structure leveraging feature relevance* approaches either act only as edge detectors and are invariant to higher-level NN weights [62]. Lack of validation also led to the failure of the first generation of ES explanations. Approaches developed by computer scientists as research tools looked reasonable to an outside observer, but were never really implemented in the real world because domain experts did not find them useful [22, pp. 1-2]. In the future, XAI approaches need to be validated and tested on users to ensure that they work as expected and to provide a reasonable basis for further research.

Both principles proposed are not completely new. As mentioned earlier, *Embedment* was at the bottom of the shift from first to second generation ES explanations. *Scientific testing* is considered for some time in the XAI debate [62]–[64]. We were again able to show why that is a good thing and how *Scientific Testing* is an important principle to consider in the future. Both principles are not just a suggestion for the future, but an important tool - together with the other principles - to analyse state-of-the-art approaches.

VI. Conclusion

Our paper analysed the current state of XAI from several perspectives. First, we provided a historical reconstruction of XAI developments. It relates AI improvements to their inevitable impact on explanatory research throughout history. The reconstruction provides the basis for a meaningful analysis of the state of XAI. It is the basis for discussing future developments of XAI for our paper and in general. Second, we proposed a new taxonomy based on this historical reconstruction and state-of-the-art XAI approaches. Our taxonomy has a flowchart inspired structure. By relating the dimensions of the taxonomy not only to the XAI development process, but also to each other, this flow architecture creates an additional layer of structure that is missing in other taxonomies. It enables a deeper analysis of XAI approaches and their benefits and shortcomings. As our taxonomy is a combination of detailed categories and broadly applicable dimensions, it is suitable for a wide variety of contexts. Third, using the historical reconstruction and our flowchart inspired taxonomy, we were then able to propose three overarching principles of XAI research: *Computing Edges*, *Dimensionality Reduction*, and *Traceability/Blaming*. These are capable of structuring the debate in a new way. Our principles are not only meant to describe the current state of XAI and propose ideas that approaches currently adhere to. We also propose two new principles that XAI approaches should follow in the future to improve their explanations. The future XAI principles are *Embedment* and *Scientific Testing*. As shown in the previous sections, the current state of XAI research requires a more precise definition of its objectives and methods (*Scientific Testing*) to meet the demands of the future. In addition, researchers should improve user interaction and context sensitivity to ensure user acceptance of explanations (*Embedment*). Our structured analysis of the history and state-of-the-art of Explainable Artificial Intelligence provides a basis for a fruitful discussion on XAI. By integrating historical perspectives with state-of-the-art approaches, our research stimulates discussion about the principles that XAI follows and should follow in the future. A more systematic approach would benefit the field in the long run, allowing for more efficient development and potentially opening it up to researchers from other disciplines in the future.

References

[1] M. Aubry and B. C. Russell, “Understanding deep features with computer-generated imagery,” in *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015.

[2] T. Dang, H. Van, H. Nguyen, V. Pham, and R. Hewett, “Deepvix,” in *Proceedings of the 11th International Conference on Advances in Information Technology*, K. Porkaew, M. Chignell, S. Fong, and B. Watanapa, Eds. New York, NY, USA: ACM, 2020, pp. 1–10.

[3] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, “Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav),” *International Conference on Machine Learning*, pp. 2668–2677, 2018. [Online]. Available: http://proceedings.mlr.press/v80/kim18d.html

[4] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec, “Interpretable & explorable approximations of black box models.” [Online]. Available: http://arxiv.org/pdf/1707.01154v1

[5] G. Liu and D. Gifford, “Visualizing feature maps in deep neural networks using deepresolve. a genomics case study,” *Proceedings of the International Conference on Machine Learning—Workshop on Visualization for Deep Learning*, no. 70, 2017. [Online]. Available: http://icmlviz.github.io/assets/papers/7.pdf

[6] J. Liu, Y. Lin, L. Jiang, J. Liu, Z. Wen, and X. Peng, *Improve Interpretability of Neural Networks via Sparse Contrastive Coding*, 2022. [Online]. Available: http://pengxi.me/wp-content/uploads/2022/10/333_paper.pdf

[7] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, “Explaining nonlinear classification decisions with deep taylor decomposition,” *Pattern Recognition*, vol. 65, pp. 211–222, 2017.

[8] P. E. Rauber, S. G. Fadel, A. X. Falcao, and A. C. Telea, “Visualizing the hidden activity of artificial neural networks,” *IEEE transactions on visualization and computer graphics*, vol. 23, no. 1, pp. 101–110, 2017.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, “”why should i trust you?”,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, B. Krishnapuram, M. Shah, A. Smola, C. Aggarwal, D. Shen, and R. Rastogi, Eds. New York, NY, USA: ACM, 2016, pp. 1135–1144.

[10] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps.” [Online]. Available: http://arxiv.org/pdf/1312.6034v2

[11] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[12] M.-Y. Kim, S. Atakishiyev, H. K. B. Babiker, N. Farruque, R. Goebel, O. R. Zaïane, M.-H. Motallebi, J. Rabelo, T. Syed, H. Yao, and P. Chun, “A multi-component framework for the analysis and design of explainable artificial intelligence,” *Machine Learning and Knowledge Extraction*, vol. 3, no. 4, pp. 900–921, 2021.

[13] M. Robnik-Šikonja and M. Bohanec, “Perturbation-based explanations of prediction models,” in *Human and Machine Learning*, ser. Human–Computer Interaction Series, J. Zhou and F. Chen, Eds. Cham: Springer International Publishing, 2018, pp. 159–175.

[14] W. Saeed and C. Omlin, “Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities,” *Knowledge-Based Systems*, vol. 263, p. 110273, 2023.

[15] G. Schwalbe and B. Finzel, “A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts,” *Data Mining and Knowledge Discovery*, 2023.

[16] C. Zednik, “Solving the black box problem: a normative framework for explainable artificial intelligence,” *Philosophy & Technology*, vol. 34, no. 2, pp. 265–288, 2021.

[17] S. T. Mueller, R. R. Hoffman, W. Clancey, A. Emrey, and G. Klein, “Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai.” [Online]. Available: https://arxiv.org/pdf/1902.01876

[18] F. Rosenblatt, “Principles of neurodynamics. perceptrons and the theory of brain mechanisms,” *American Journal of Psychology*, vol. 76, p. 705, 1963.

[19] J. Lighthill, “Artificial intelligence: A general survey.” [Online]. Available: http://www.chilton-computing.org.uk/inf/literature/reports/lighthill_report/p001.htm

[20] B. G. Buchanan and E. A. Feigenbaum, “Dendral and meta-dendral: Their applications dimension,” *Artificial Intelligence*, vol. 11, no. 1-2, pp. 5–24, 1978.

[21] E. H. Shortliffe, *Computer-based medical consultations: MYCIN*, ser. Artificial intelligence series. New York: North-Holland, 1976, vol. 2.

[22] J. D. Moore and W. R. Swartout, *Explanation in expert systems: A survey*, 1988. [Online]. Available: https://scholar.google.de/citations?user=iqlmqryaaaaj&hl=de&oi=sra

[23] J. Durkin, “Expert systems: a view of the field,” *IEEE Expert*, vol. 11, no. 2, pp. 56–63, 1996.

[24] W. J. Clancey, “Guidon,” in *The Handbook of Artificial Inielligence*, A. Barr and E. A. Feigenbaum, Eds. William Kaufmann, Inc., 1982, pp. 8–15. [Online]. Available: https://billclancey.name/GUIDON-Clancey-CBI1982.pdf

[25] W. R. Swartout, “Xplain: a system for creating and explaining expert consulting programs,” *Artificial Intelligence*, vol. 21, no. 3, pp. 285–325, 1983. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370283800149

[26] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual explanations without opening the black box: Automated decisions and the gdpr,” *Harvard Journal of Law & Technology*, vol. 31, no. 2, 2018. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3063289

[27] J. L. Weiner, “Blah, a system which explains its reasoning,” *Artificial Intelligence*, vol. 15, no. 1-2, pp. 19–48, 1980.

[28] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[29] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 618–626.

[30] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, and K. Baum, “What do we want from explainable artificial intelligence (xai)? – a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research,” *Artificial Intelligence*, vol. 296, no. 3, p. 103473, 2021. [Online]. Available: http://arxiv.org/pdf/2102.07817v1

[31] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[32] Z. C. Lipton, “The mythos of model interpretability,” *Communications of the ACM*, vol. 61, no. 10, pp. 36–43, 2018.

[33] L. Li, B. Wang, M. Verma, Y. Nakashima, R. Kawasaki, and H. Nagahara, “Scouter: Slot attention-based classifier for explainable image recognition,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Ed., 2021, pp. 1046–1055.

[34] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *Computer Vision – ECCV 2014*, ser. Lecture Notes in Computer Science, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, vol. 8689, pp. 818–833.

[35] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in neural information processing systems 30*, U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds. Red Hook, NY: Curran Associates Inc, 2017, vol. 30. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978\632d76c43dfd28b67767-Paper.pdf

[36] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Advances in neural information processing systems*, D. D. Lee, U. von Luxburg, R. Garnett, M. Sugiyama, and I. Guyon, Eds., vol. 29. Red Hook, NY: Curran Associates Inc, 2016. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2016/file/231141b34c82\aa95e48810a9d1b33a79-Paper.pdf

[37] S. Barratt, "Interpnet: Neural introspection for interpretable deep learning."

[38] A. Ali, T. Schnake, O. Eberle, G. Montavon, K.-R. Müller, and L. Wolf, "Xai for transformers: Better explanations through conservative propagation," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, Le Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022, pp. 435–451. [Online]. Available: https://proceedings.mlr.press/v162/ali22a.html

[39] D. Kazhdan, B. Dimanov, M. Jamnik, and P. Liò, "Meme: Generating rnn model explanations via model extraction." [Online]. Available: http://arxiv.org/pdf/2012.06954v1

[40] T. Speith, "A review of taxonomies of explainable artificial intelligence (xai) methods," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. ACM Digital Library. New York,NY,United States: Association for Computing Machinery, 2022, pp. 2239–2250.

[41] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net." [Online]. Available: http://arxiv.org/pdf/1412.6806v3

[42] A. Adhikari, D. M. J. TaxTax, R. Satta, and M. Faeth, "Leafage: Example-based and feature importance-based explanations for black-box ml models," in *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, IEEE, Ed., 2019, pp. 1–7. [Online]. Available: https://ieeexplore.ieee.org/document/8858846

[43] G. Alain and Y. Bengio, "Understanding intermediate layers using linear classifier probes." [Online]. Available: https://arxiv.org/abs/1610.01644

[44] H. Tan and H. Kotthaus, "Surrogate model-based explainability methods for point cloud nns," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, IEEE, Ed., 2022, pp. 2239–2248.

[45] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, p. e0130140, 2015. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130140

[46] O. Bastani, C. Kim, and H. Bastani, "Interpreting blackbox models via model extraction."

[47] U. Ehsan, B. Harrison, L. Chan, and M. O. Riedl, "Rationalization: A neural machine translation approach to generating natural language explanations," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, J. Furman, G. Marchant, H. Price, and F. Rossi, Eds. New York, NY, USA: ACM, 2018, pp. 81–87.

[48] J. Ribeiro, L. Cardoso, R. Silva, V. Cirilo, N. Carneiro, and R. Alves, "Global explanation of tree-ensembles models based on item response theory." [Online]. Available: http://arxiv.org/pdf/2210.09933v1

[49] R. Tomsett, D. Braines, D. Harborne, A. Preece, and S. Chakraborty, "Interpretable to whom? a role-based model for analyzing interpretable machine learning systems," *ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)*, 2018. [Online]. Available: http://arxiv.org/pdf/1806.07552

[50] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[51] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, 2019.

[52] A. M. Antoniadi, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B. A. Becker, and C. Mooney, "Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: A systematic review," *Applied Sciences*, vol. 11, no. 11, p. 5088, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/11/5088

[53] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 3–19.

[54] O. Buchholz, "A means-end account of explainable artificial intelligence," *Synthese*, vol. 202, no. 2, pp. 1–23, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s11229-023-04260-w

[55] C. G. Hempel, "Laws and their role in scientific explanation," in *The philosophy of science*, ser. A Bradford book, R. Boyd, P. Gasper, and J. D. Trout, Eds. Cambridge, Mass.: MIT Press, 1991, pp. 299–315.

[56] R. Neches, W. Swartout, and J. D. Moore, "Explainable (and maintainable) expert systems," in *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, A. Joshi, Ed. Los Altos, Calif.: Morgan Kaufmann, 1985. [Online]. Available: https://api.semanticscholar.org/CorpusID:1376515

[57] K. Ford, J. Coffey, A. Cañas, E. Andrews, and C. Turner, "Diagnosis and explanation by a nuclear cardiology expert system," *International Journal of Expert Systems*, vol. 9, p. 499, 1996.

[58] T. Spinner, U. Schlegel, H. Schafer, and M. El-Assady, "explainer: A visual analytics framework for interactive and explainable machine learning," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 1064–1074, 2020.

[59] A. Hedström, L. Weber, D. Krakowczyk, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, and M. M. C. Höhne, "Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond," *Journal of Machine Learning Research*, vol. 24, no. 34, pp. 1–11, 2023. [Online]. Available: http://jmlr.org/papers/v24/22-0142.html

[60] F. Stieler, F. Rabe, and B. Bauer, "Towards domain-specific explainable ai: Model interpretation of a skin image classifier using a human approach," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2021, pp. 1802–1809. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021W/ISIC/html/Stieler_\Towards_Domain-Specific_Explainable_AI_Model_Interpretation_of_a_Skin_Image_CVPRW_2021_paper.html

[61] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in ai," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, 2019, pp. 279–288.

[62] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Advances in neural information processing systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Red Hook, NY: Curran Associates Inc, 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/294a8ed24b1a\d22ec2e7efea049b8737-Paper.pdf

[63] A. Alqaraawi, M. Schuessler, P. Weiß, E. Costanza, and N. Berthouze, "Evaluating saliency map explanations for convolutional neural networks," in *Proceedings of the 25th International Conference on Intelligent User Interfaces*, F. Paternò, N. Oliver, C. Conati, L. D. Spano, and N. Tintarev, Eds. New York, NY, USA: ACM, 2020, pp. 275–285.

[64] A. H. A. Rahnama, J. Bütepage, P. Geurts, and H. Boström, "Can local explanation techniques explain linear additive models?" *Data Mining and Knowledge Discovery*, vol. 38, no. 1, pp. 237–280, 2024.