Unifying VXAI: A Systematic Review and Framework for the Evaluation of Explainable AI

David Dembinsky *david.dembinsky@dfki.de*
German Research Center for Artificial Intelligence

Adriano Lucieri *adriano.lucieri@dfki.de*
German Research Center for Artificial Intelligence

Stanislav Frolov *stanislav.frolov@dfki.de*
German Research Center for Artificial Intelligence

Hiba Najjar *hiba.najjar@dfki.de*
German Research Center for Artificial Intelligence

Ko Watanabe *ko.watanabe@dfki.de*
German Research Center for Artificial Intelligence

Andreas Dengel *andreas.dengel@dfki.de*
German Research Center for Artificial Intelligence

Abstract

Modern AI systems frequently rely on opaque black-box models, most notably Deep Neural Networks, whose performance stems from complex architectures with millions of learned parameters. While powerful, their complexity poses a major challenge to trustworthiness, particularly due to a lack of transparency. Explainable AI (XAI) addresses this issue by providing human-understandable explanations of model behavior. However, to ensure their usefulness and trustworthiness, such explanations must be rigorously evaluated. Despite the growing number of XAI methods, the field lacks standardized evaluation protocols and consensus on appropriate metrics. To address this gap, we conduct a systematic literature review following the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)* guidelines and introduce a unified framework for the *eValuation of XAI (VXAI)*. We identify 362 relevant publications and aggregate their contributions into 41 functionally similar metric groups. In addition, we propose a three-dimensional categorization scheme spanning explanation type, evaluation contextuality, and explanation quality desiderata. Our framework provides the most comprehensive and structured overview of VXAI to date. It supports systematic metric selection, promotes comparability across methods, and offers a flexible foundation for future extensions.

1 Introduction

Explainable AI (XAI) is a research area of growing interest to both AI researchers and practitioners. It aims to alleviate the black-box issue of current deep-learning models, which can reach stunning performances at the expense of their interpretability (Vilone & Longo, 2021). Government-affiliated initiatives, such as the European Union High-Level Expert Group on AI (2019), the U.S. National

Institute of Standards and Technology (2023), and the DARPA initiative (Gunning & Aha, 2019), identified XAI as a crucial part of Trustworthy AI. Especially as it helps AI systems in serving the "right to explain" its decisions (Goodman & Flaxman, 2017) and fosters user trust through understanding of the system (Morandini et al., 2023). XAI already plays a fundamental role in making high-stakes AI systems more trustworthy (Saarela & Podgorelec, 2024; Xua & Yang, 2024), with broad applications in areas such as healthcare, finance, autonomous driving, natural disaster detection, energy management, military and remote sensing (Adadi & Berrada, 2018; Markus et al., 2021; Saraswat et al., 2022; Kadir et al., 2023; Hosain et al., 2024; Höhl et al., 2024). Furthermore, explainability is used to help with other dimensions of trustworthiness like privacy, robustness, or fairness (Doshi-Velez & Kim, 2017; Yang et al., 2019; Arrieta et al., 2020; Das & Rad, 2020; Markus et al., 2021; Rawal et al., 2021; Agarwal et al., 2022b).

However, XAI is not a silver bullet. Van der Waa et al. (2021) point out, that humans tend to trust predictions more readily when an explanation is provided, often without carefully examining the explanation itself. This lack of critical scrutiny can lead to unwarranted trust, especially when decisions are taken based on incorrect or misleading explanations (Eiband et al., 2019; Jesus et al., 2021). To make matters worse, different XAI algorithms may result in conflicting explanations for the same model and sample (Krishna et al., 2022). Therefore, simply providing *any* explanation is not sufficient, but it is important to assess the quality of the explanation at hand (Sovrano et al., 2021). Unfortunately, while there is a plethora of XAI methods, evaluation of explanations is still an immature research area (Ribera & Lapedriza, 2019), with many studies relying on the notion of a good explanation as "You'll know it when you see it", providing anecdotal evidence (i.e. small-scale qualitative validation) (Doshi-Velez & Kim, 2017; Nauta et al., 2023; Saarela & Podgorelec, 2024). Especially in computer vision tasks, evaluation through qualitative inspection of a few examples can be appealing (Ibrahim & Shafiq, 2023). However, unstructured qualitative examination leads to highly subjective results, as humans struggle at judging the value of XAI explanations (Adebayo et al., 2018; Buçinca et al., 2020; Hase & Bansal, 2020). In addition, such evaluations risk cherry-picking favorable examples and offer no reliable foundation for comparing different explanation methods across studies or practitioners. For the same given explanation, human ratings vary depending on both the task itself (Franklin & Lagnado, 2022) and the participant's cultural background (Peters & Carman, 2024). Evaluation is further complicated by the lack of ground-truth for the explanations, as this would require knowledge about the model's internal reasoning process (Samek et al., 2019; Markus et al., 2021; Samek et al., 2021; Bommer et al., 2024; Ortigossa et al., 2024).

Because evaluation is still not performed consistently and seldom systematically (Adadi & Berrada, 2018; Lipton, 2018; Payrovnaziri et al., 2020; Messina et al., 2022; Lopes et al., 2022; De Camargo et al., 2023; Kadir et al., 2023; Nauta et al., 2023; Mohamed et al., 2024; Naveed et al., 2024; Saarela & Podgorelec, 2024; Salih et al., 2024a), the community frequently calls to develop comprehensive and unified evaluation standards (Pinto & Paquette, 2024; Saarela & Podgorelec, 2024; Xua & Yang, 2024). A central motivation behind such efforts is to enable the comparison of explanations and asses whether explainability is achieved (Markus et al., 2021; Zhou et al., 2021)

One of the most prevalent taxonomies reported in the literature (Vilone & Longo, 2021; Zhou et al., 2021; Elkhawaga et al., 2023), and illustrated in Figure 1, is the distinction proposed by Doshi-Velez & Kim (2017) between human-grounded and functionality-grounded evaluation methods. The former includes qualitative and quantitative evaluations by laypeople and experts, while the latter consists of (semi-)automatic metrics.

Since explanations are meant to aid humans, human-grounded evaluation remains the gold standard to assess their effectiveness in assisting humans (Doshi-Velez & Kim, 2017; Gunning & Aha, 2019; Miller et al., 2017). However, the faithfulness (i.e., technical correctness) of an explanation and the plausibility to humans do not necessarily correlate (Wiegreffe & Pinter, 2019; Jacovi & Goldberg, 2020; Atanasova, 2024a). Therefore, human-grounded evaluation of an explanation's comprehensibility should be distinguished from functionality-grounded evaluation of its faithfulness (Nauta et al., 2023). Especially, humans can't confidently attribute whether an unexpected explanans (i.e., the information provided to explain a decision[1]) is caused by a faulty explanation (process[1]) or a flawed black-box model (Robnik-Šikonja & Bohanec, 2018; Zhang et al., 2019a); see Figure 2 for an illustration. In

[1]The exact definitions of explanandum, explanation, and explanans are provided at the end of Section 1.

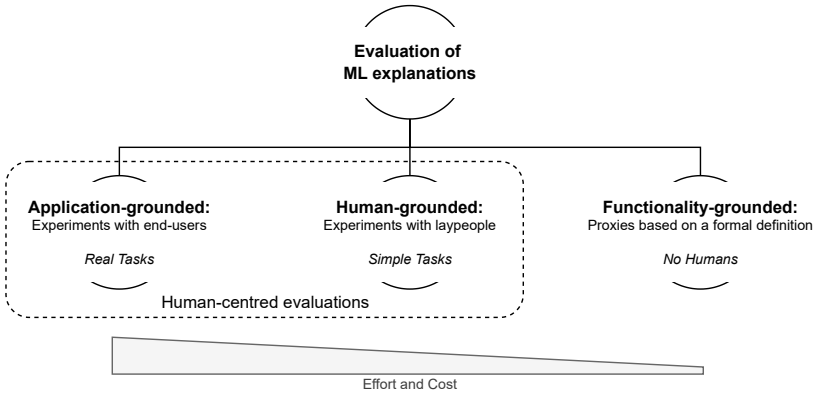

Figure 1: The classification of XAI evaluation into human-grounded and functionality-grounded evaluation, adapted from the classification framework by Doshi-Velez & Kim (2017) and its visualization by Zhou et al. (2021).

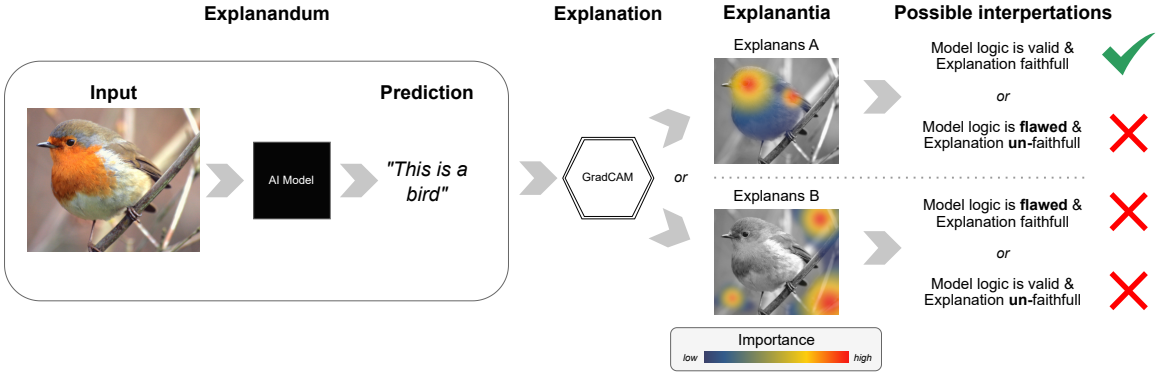

Figure 2: The heatmaps show two alternative example explanantia[1], indicating which input regions were deemed decisive for the model's decision (the explanandum[1]). A qualitative inspection allows for multiple interpretations, as it is unclear whether a) both the model and the explanation process (explanation[1]) are correct or flawed (top), or b) one is correct and the other failed (bottom).

both cases, the consequences can be severe, either reducing trust in a well-functioning model or, more critically, reinforcing trust in a flawed one. Further, human evaluation, especially through the system's developers, is prone to confirmation bias (Doshi-Velez & Kim, 2017; Lipton, 2018)

There exist a number of surveys and guidelines that address human-centered evaluations (Hoffman et al., 2018; Miller, 2019; Chromik & Schuessler, 2020; Holzinger et al., 2020; Franklin & Lagnado, 2022; Hsiao et al., 2021; Jesus et al., 2021; Langer et al., 2021; Mohseni et al., 2021; van der Waa et al., 2021; Silva et al., 2023). Unfortunately, the range of reviews dedicated to functionality-grounded evaluation is still limited. This is despite the advantage of offering objective, quantitative metrics without requiring human experiments, which can save both time and cost (Doshi-Velez & Kim, 2017; Samek et al., 2019; Zhou et al., 2021). Most existing studies are narrow and restricted to a specific application domain (Giuste et al., 2022; Arreche et al., 2024), including cybersecurity (Pawlicki et al., 2024), medical image classification (Patrício et al., 2023; Chaddad et al., 2024), electronic health record data (Payrovnaziri et al., 2020), data and knowledge engineering (Li et al., 2020b), or time-series classification (Theissler et al., 2022). Others focus on particular XAI approaches, such as visual explanations in CNNs (Mohamed et al., 2022) or instance-based explanations (Bayrak & Bach, 2024). Moreover, many surveys dedicate only limited attention to evaluation metrics, with their primary focus placed on the XAI methods themselves (Carvalho et al., 2019; Ding et al., 2022; Minh et al., 2022; Mohamed et al., 2022; Ali et al., 2023; Clement et al., 2023; Patrício et al., 2023; Chaddad et al., 2024; Gongane et al., 2024; Xua & Yang, 2024). By contrast, this review focuses exclusively on functionality-grounded evaluation across domains and is applicable to a wide range of XAI approaches.

Contributions

Despite the growing number of proposed metrics, a comprehensive and unified framework for functionality-grounded evaluation is still missing. Further, the inconsistent use of terms such as interpretability, comprehensibility, understandability, transparency, and explainability (Koh & Liang, 2017; Guidotti et al., 2018; Arrieta et al., 2020; Markus et al., 2021) hampers conceptual clarity and comparability across approaches. To address this gap, we introduce a framework called **eValuation of Explainable Artificial Intelligence (VXAI)**, aimed at unifying functionality-grounded evaluation for XAI.
Our contributions are as follows:

- We perform a systematic literature review based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines by Page et al. (2021), identifying 362 relevant publications that introduce or utilize evaluation metrics.
- We aggregate these into 41 functionally similar metric groups, capturing common methodological patterns across the literature.
- We propose a three-dimensional categorization scheme consisting of desiderata, explanation type, and evaluation contextuality, and use it to organize the identified metrics.
- To our knowledge, this results in the most comprehensive and unified VXAI framework to date and provides an extensible foundation for future research.

The remainder of this review is structured as follows: In Section 2, we first present related studies on the topic of VXAI to motivate the need for this systematic review, before introducing the desiderata of XAI in Section 3. Further, Section 4 introduces our research method used to search for and identify relevant metrics. The categorization scheme and results are presented in Section 5, where we introduce our categorization framework (Subsection 5.1) and summarize the identified metrics (Subsection 5.2), complemented by a visual overview in Figure 6. A deeper discussion of these findings is provided in Section 6, while comprehensive descriptions of the metrics alongside references are listed in Appendix B. We conclude the review in Section 7, discussing the results and future paths for the area of VXAI.

Terminology

To avoid ambiguous language, throughout the paper we stick to the terminology of the XAI Handbook by Palacio et al. (2021): The goal of XAI is to facilitate understanding by providing insights into an *explanandum* ("What is to be explained"), usually a model or a model's decision. To accomplish this, we leverage an *explanation*, which is the process of getting insight into the explanandum. The resulting output of this process is the *explanans*, which provides the user with information about the model's inner workings. In a mathematical sense, the explanation can be viewed as a function that maps an explanandum to an explanans. For example, the explanandum could be a CNN's classification of a given input image. The explanation might be an algorithm such as GradCAM (Selvaraju et al., 2017), and the resulting heatmap is the explanans, which highlights important features. We use the Latin plural forms explananda (explanandum) and explanantia (explanans) throughout. When we refer to VXAI, we include both the evaluation of the method (explanation) and its output (explanans), since most evaluation metrics necessarily assess the quality of explanations through the quality of their generated outputs.

As defined by the XAI Handbook, *interpretation* (or *interpretability*) refers to the subsequent assignment of meaning to an explanation. It describes the process through which a human infers knowledge about the explanandum using the explanans. This step significantly influences the success of the explanation and also depends on the receiving human's (the *explainee*'s) mental model.

Terminology

explanandum (pl. **explananda**): What is to be explained, i.e. a model and its prediction.
explanation (pl. **explanations**): The process of explaining, i.e. the XAI algorithm.
explanans (pl. **explanantia**): The explaining information, i.e. the output of an explanation.

2 Related Work

Although the field of XAI has gained popularity over the past years, there is still no extensive and unified evaluation framework for XAI metrics. Various surveys have explored XAI and VXAI from different angles, ranging from human-grounded evaluation to technical metrics. Table 1 gives an overview over 30 such XAI reviews from the past years.

While evaluation of XAI is frequently given less attention in XAI surveys, 23 of these reviews directly focus on the topic of VXAI. Besides functionality-grounded evaluation, a second school of thought is concerned with human-grounded evaluation of explanations through qualitative expert evaluations or quantitative user studies, with representative surveys for this domain available as well (Sokol & Flach, 2020; Rawal et al., 2021; Naveed et al., 2024). Nevertheless, a considerable number of 19 reports focus specifically on the topic of functionality-grounded evaluation. Unfortunately, most of these surveys focus on a subset of well-known metrics, whereas only 14 surveys gathered VXAI metrics in a systematic or semi-systematic literature review. Further, numerous of the referenced reviews either lack an extensive list of desiderata and focus only on a subset of them, or limit their research to specific types of explanations[2] or application domains.

There are five reviews, that we consider most similar to this work, as they present systematic functionality-grounded VXAI surveys. Le et al. (2023) and Nauta et al. (2023) both categorize the identified metrics based on a scheme of 12 properties, namely the Co-12 framework, which we discuss in more detail in Section 3. However, Le et al. (2023) restrict their analysis to metrics available through public XAI or VXAI toolkits (e.g., Quantus (Hedström et al., 2023)) and do not report on metrics introduced in the literature but not implemented in such libraries. Although there is some overlap with the study by Nauta et al. (2023), particularly in the inclusion of some identical metrics, their review also incorporates studies that merely apply VXAI metrics rather than introducing them. In contrast, our work provides detailed descriptions and categorizations of each identified metric. The review by Kadir et al. (2023) covers a broad range of domains and explanation types[2] and reports a wide variety of metrics. It also groups several metrics by method, a strategy shared by our work. However, it does not adopt a categorization scheme based on the desiderata fulfilled by individual metrics. Notably, the recent reviews from Bayrak & Bach (2024) and Pawlicki et al. (2024) report a high number of individual metrics for VXAI. However, both limit the scope of their review considerably, either in terms of application domain (Pawlicki et al., 2024) or explanation type[2] (Bayrak & Bach, 2024). In contrast, our work includes all metrics reported to date and introduces a categorization scheme based on explicitly defined desiderata (see Section 3). Finally, many of the metrics we identified were introduced only recently, underlining the need for this more recent literature review.

While previous reviews report between 10 and 90 individual metrics, our work introduces a unified structure by aggregating over 360 individual metrics into 41 conceptually related groups. This enables clearer comparison and interpretation across metrics. Unlike most surveys, we do not limit our analysis to specific explanation types or application domains, ensuring broader applicability across the XAI landscape.

3 Desiderata of XAI

A well-founded evaluation of XAI methods requires clearly defined criteria for what constitutes a good explanation. To establish such criteria, we must first reflect on the role of explanations in the context of XAI. According to the definition in the XAI Handbook (Palacio et al., 2021), explaining a model and its behavior is a two-stage process: first, factual information about the model's decision process is generated (the explanans); this is then interpreted by the human user. The first stage can be evaluated using technical criteria that assess whether the model's reasoning has been captured truthfully and reliably. The second stage depends on the interpretability of the explanation, which can be assessed using general cognitive principles, even in the absence of a specific user model.

To capture the multifaceted nature of explanation quality, a number of desiderata have been proposed in the literature. We interpret these as functionality-grounded expectations that reflect the demands

[2]The *explanation type* refers to both the design of the explanation algorithm and, consequently, the nature of the resulting explanans, as introduced in Subsection 5.1.2.

Work	Primary VXAI Focus	Primarily Functionality-Grounded VXAI	(Semi-)Systematic Review	Date ↓	Desiderata	Limited to	Reported Metrics
This work	✓	✓	✓	Jan 2025	Parsimony, Plausibility, Coverage, Fidelity, Continuity, Consistency, Efficiency		41 metrics from 362 sources
Klein et al.	✓	✓		Jan 2025	Faithfulness, Robustness, Complexity	Feature Attributions; Computer Vision	20 metrics
Pawlicki et al.	✓	✓	✓	Oct 2024		Cybersecurity	86 metrics
Awal & Roy	✓	✓		Jun 2024	Reliability, Consistency	Rule Explanations	6 metrics
Bayrak & Bach	✓	✓	✓	Apr 2024		Counterfactuals	66 metrics
Bommer et al.	✓	✓		Mar 2024	Robustness, Faithfulness, Complexity, Localization, Randomization	Climate Science	10 metrics
Li et al.	✓	✓		Dec 2023	Faithfulness	Feature Attributions	6 metrics
Alangari et al.	✓			Aug 2023	Correctness, Comprehensibility, Stability		59 metrics
Le et al.	✓	✓	✓	Aug 2023	Co-12†		86 metrics from 17 toolkits
Salih et al.	✓		✓	Aug 2023		Cardiology	27 metrics
Kadir et al.	✓	✓	✓	Jul 2023			80 metrics
Hedström et al.	✓	✓		Apr 2023	Faithfulness, Robustness, Localization, Complexity, Axiomatic, Randomization	Feature Attribution	27 metrics
Schwalbe & Finzel			✓	Jan 2023			11 metrics (already grouped)
Agarwal et al.	✓	✓		Nov 2022	Faithfulness, Stability	Feature Attributions	11 metrics
Coroama & Groza	✓			Nov 2022			26 metrics
Verma et al.		✓		Nov 2022		Counterfactuals	9 metrics (already grouped)
Belaid et al.	✓	✓		Oct 2022	Fidelity, Fragility, Stability, Simplicity, Stress, Other	Feature Attributions	22 metrics
Cugny et al.	✓	✓		Oct 2022			6 metrics
Lopes et al.	✓		✓	Aug 2022	Fidelity (Completeness, Soundness), Interpretability, Broadness, Simplicity, Clarity)		43 metrics
Yuan et al.		✓	✓	Jul 2022	Fidelity, Sparsity, Stability, Accuracy	Graph Neural Networks	7 metrics
Löfström et al.	✓		✓	Mar 2022			10 metrics
Vilone & Longo	✓		✓	Dec 2021			36 metrics
Bodria et al.		✓	✓	Nov 2021			6 metrics
Sovrano et al.	✓			Oct 2021	Similarity, Exactness, Fruitfulness		22 metrics
Ras et al.				Sep 2021			*13 sources in text* (metrics not listed)
Mohseni et al.			✓	Aug 2021	Fidelity, Trustworthiness		15 metrics
Yeh & Ravikumar	✓	✓		Jun 2021			7 metrics
Nauta et al.	✓	✓	✓	May 2021	Co-12†		28 metrics (already grouped)
Zhou et al.	✓	✓		Jan 2021	Fidelity (Completeness, Soundness), Interpretability, Broadness, Simplicity, Clarity)		17 metrics
Samek & Müller				Sep 2019			*16 sources in text* (metrics not listed)
Yang et al.	✓	✓	✓	Aug 2019	Generalizability, Fidelity, Persuasibility		*40 sources in text* (metrics not listed)

†Co-12: Correctness, Output-Completeness, Consistency, Continuity, Contrastivity, Covariate Complexity, Compactness, Composition, Confidence, Context, Coherence, Controllability

Table 1: Overview of recent XAI reviews, sorted by date. The table indicates whether each survey primarily focused on evaluation metrics, whether it reported mainly functionality-grounded metrics, and whether a (semi-)structured review was conducted. The date refers to the earliest available point in the article's timeline; either the database query, submission, or publication, depending on what was reported. For each survey, we also report the desiderata used to classify the metrics and any limitations regarding explanation type or application domain. Note that not all surveys systematically listed their assessed metrics, so the reported metric count may vary depending on the method of extraction.

of both stages of the explanation process. In this section, we first provide an overview of existing desiderata proposed in prior work. We then introduce a unified framework that systematically describes the requirements for ensuring technical soundness and for bridging the interpretation gap.

3.1 Common Formulation of Desiderata

Several XAI surveys report that there is no ubiquitous consensus on appropriate desiderata, with some of the categories related to goals pursued *through* XAI, rather than standalone desiderata of XAI, e.g., Trustworthiness, Acceptance, or Fairness (Doshi-Velez & Kim, 2017; Langer et al., 2021; Vilone & Longo, 2021; Elkhawaga et al., 2023). Hence, we conduct a scoping review, reporting the main desiderata used by different authors and analyzing the similarities as well as differences in their formulations. For the sake of brevity, we exclude some of the papers listed in Table 1, as the missing ones either overlap considerably (e.g. Awal & Roy (2024) and Klein et al. (2024)), rely on a different notion of desiderata (e.g., Sovrano et al. (2021)), or use no desiderata at all. We will first present these frameworks using the authors' original terminology before introducing our own categorization scheme.

The famous **Co-12** properties, introduced by Nauta et al. (2023) and reused by Le et al. (2023), constitute one of the most extensive existing frameworks for categorizing XAI metrics. They group the properties along three different dimensions: Content (*Correctness, Completeness, Consistency, Continuity, Contrastivity, Covariate Complexity*), Presentation (*Compactness, Composition, Confidence*), and User (*Context, Coherence, Controllability*). While the first dimension focuses on the information contained in the explanans, the second and third dimensions address the way the information is conveyed. Although some of these human-centered properties can be measured through proxies, others may mainly be evaluated through human-grounded evaluation.

Zhou et al. (2021), based on the taxonomy of Markus et al. (2021), define *Interpretability* and *Fidelity* as the two major components of explainability. The former is concerned with providing understandable explanantia and includes the properties of *Clarity*, *Broadness*, and *Parsimony*. Fidelity, on the other hand, refers to how accurately an explanans reflects the model's behavior, and consists of the properties of *Completeness* and *Soundness*.

The framework proposed by Robnik-Šikonja & Bohanec (2018), which was adopted by Carvalho et al. (2019) and Molnar (2020), differentiates between properties of explanations and individual explanantia. Investigating the properties of methods (i.e. explanations), they consider *Translucency*, *Portability*, and *Algorithmic Complexity*, which can all be interpreted as desiderata, while *Expressive Power* is a descriptive property. The properties of individual explanantia include *Comprehensibility*, *Importance*, *Representativeness*, *Fidelity*, and *Stability*. However, their categorization encompasses further properties, which we do not consider as proper desiderata of XAI: *Accuracy*, *Novelty*, *Certainty*, and *Consistency*. Accuracy is a measurement of the underlying black-box model, while Novelty and Certainty are rather explanantia themselves, than properties of general explanations. Further, Consistency between different black-box models is not necessarily a useful measure, as different models may derive similar predictions based on different reasoning (see Rashomon Effect (Breiman, 2001; Leventi-Peetz & Weber, 2022)).

The famous XAI review by Guidotti et al. (2018), inspired by earlier works such as Andrews et al. (1995) and Johansson et al. (2004), reports three less fine-grained desiderata: *Interpretability*, *Fidelity*, and *Accuracy*. Similar to previously discussed reviews, Interpretability describes human understandability, while Fidelity measures how well the explanans imitates the black box, and Accuracy focuses on predictive performance, which is outside the scope of XAI in our context. Additionally, *Consistency* is introduced by Andrews et al. (1995), expecting reproducible explanations, while Johansson et al. (2004) emphasize the explanation's algorithmic *Scalability* and *Generality*.

Alvarez-Melis & Jaakkola (2018b), Jesus et al. (2021), and Alangari et al. (2023a) all report a similar set of desiderata. The understandability of explanations is measured in terms such as *Interpretability*, while *Faithfulness* and the corresponding desiderata give insight into how truthful the explanation is to the underlying black-box model. All three works further report the *Stability* of explanations as a desired property, assessing whether explanantia on similar inputs are similar.

In their Quantus toolkit, Hedström et al. (2023) (and the follow-up study by Bommer et al. (2024) as well), categorize their metrics partly through desiderata, namely *Faithfulness*, *Robustness*, and

Complexity. Simultaneously, part of their metrics are grouped by their conceptual similarity, including *Localization*, *Randomization*, and *Axiomatic*.

Finally, the Compare-xAI benchmark by Belaid et al. (2022) organizes functional tests into six categories, namely *Fidelity*, the robustness-related *Stability* and *Fragility*, the interpretability desideratum *Simplicity*, and the explanation-methods-focused *Stress* and *Portability* (which they integrate under "Other").

While many existing frameworks overlap conceptually, a unified and practically usable categorization scheme for VXAI metrics is still lacking. This requires a structured set of desiderata that defines what makes a good explanation and supports consistent classification of metrics. Prior work often enforces a rigid one-to-one mapping between metrics and desiderata; in contrast, we decouple these dimensions, defining a set of mostly independent desiderata to which each metric may contribute individually or jointly. Lightweight frameworks tend to omit critical aspects of explanation quality, while broader ones sometimes include goals that are not intrinsic to the explanation itself (e.g., accuracy). We restrict our scope to properties that reflect the explanation rather than the underlying model and clarify excluded cases after presenting our set. Although all desiderata rely on proxies, we limit ourselves to properties that are quantifiable in principle. Highly abstract or vague notions lacking empirical grounding are omitted. Lastly, our framework is designed to be extensible, allowing the integration of future desiderata as the field evolves.

3.2 Proposed Framework of Desiderata

Building on the desiderata frameworks established above and our findings on VXAI metrics, we propose a set of seven desiderata to serve as a categorization scheme for VXAI.

Our goal is to offer a principled yet practical structure that enables consistent classification while avoiding the limitations of prior frameworks. These are either too narrow to accommodate relevant metrics or too broad and include properties beyond explainability. While properties such as fairness are often measured using XAI methods, we consider them beyond the scope of VXAI, because they assess the model's behavior rather than the explanation itself.

Building on the two-stage view of explaining described by Palacio et al. (2021) (i.e., presenting factual information followed by human interpretation), we define two complementary dimensions of explanation quality. The **Technical** (T) dimension comprises desiderata that assess the factual correctness, robustness, and reliability of the explanation, ensuring that it faithfully reflects the model's reasoning. In contrast, the **Interpretability** (I) dimension captures how the explanation is conveyed and how accessible, intuitive, and useful it is to a general-purpose user. This separation is aligned with existing frameworks such as the Co-12 properties (Nauta et al., 2023) and the taxonomy by Zhou et al. (2021). The desiderata are designed to be as independent from each other as possible, allowing for reliable quantification of different aspects relevant to trustworthy XAI.

We present our categorization scheme and its relation to other frameworks in Figure 3 and introduce them in more detail in the following paragraphs. In total, we define seven desiderata, two associated with Interpretability, and five belonging to the Technical dimension:

(I) **Parsimony**: The explanation should keep the explanans concise to support interpretability.
(I) **Plausibility**: The explanation should shape the explanans to align with human expectations.
(T) **Coverage**: The explanation should provide an explanans for every explanandum.
(T) **Fidelity**: The explanation should make the explanans reflect the model's true reasoning.
(T) **Continuity**: The explanation should ensure that similar explananda yield similar explanantia.
(T) **Consistency**: The explanation should produce stable explanantia across repeated evaluations.
(T) **Efficiency**: The explanation should compute the explanans efficiently and broadly.

3.2.1 Parsimony

The explanation should keep the explanans concise to support interpretability.