applied sciences

MDPI

Review

XAI Systems Evaluation: A Review of Human and Computer-Centred Methods

Pedro Lopes [1,*], Eduardo Silva [1,*], Cristiana Braga [1], Tiago Oliveira [2] and Luís Rosado [1]

[1] Fraunhofer Portugal AICOS, Rua Alfredo Allen 455/461, 4200-135 Porto, Portugal
[2] First Solutions—Sistemas de Informação S.A., 4450-102 Matosinhos, Portugal
* Correspondence: pedro.lopes@fraunhofer.pt (P.L.); eduardo.silva@fraunhofer.pt (E.S.)

Abstract: The lack of transparency of powerful Machine Learning systems paired with their growth in popularity over the last decade led to the emergence of the eXplainable Artificial Intelligence (XAI) field. Instead of focusing solely on obtaining highly performing models, researchers also develop explanation techniques that help better understand the system's reasoning for a particular output. An explainable system can be designed, developed, and evaluated from different perspectives, which enables researchers from different disciplines to work together on this topic. However, the multidisciplinary nature of XAI systems creates new challenges for condensing and structuring adequate methodologies to design and evaluate such systems. This paper presents a survey of Human-centred and Computer-centred methods to evaluate XAI systems. We propose a new taxonomy to categorize XAI evaluation methods more clearly and intuitively. This categorization gathers knowledge from different disciplines and organizes the evaluation methods according to a set of categories that represent key properties of XAI systems. Possible ways to use the proposed taxonomy in the design and evaluation of XAI systems are also discussed, alongside with some concluding remarks and future directions of research.

Keywords: explainable artificial intelligence; evaluation methods; human-centred; computer-centred; literature review

Citation: Lopes, P.; Silva, E.; Braga, C.; Oliveira, T.; Rosado, L. XAI Systems Evaluation: A Review of Human and Computer-Centred Methods. *Appl. Sci.* **2022**, *12*, 9423. https://doi.org/10.3390/app12199423

Academic Editors: María Paz Sesmero Lorente, Plamen Angelov and Jose Antonio Iglesias Martinez,

Received: 31 August 2022
Accepted: 15 September 2022
Published: 20 September 2022

1. Introduction

Machine Learning (ML) systems have significantly grown in popularity in the last decade. They are currently being used in several fields with increasing task-solving capabilities, from a human's everyday life (e.g., language translation) to decision-making in high-stake domains (e.g., clinical decision support) . However, most ML systems are labeled as "black-box" models because their underlying structures are complex, nonlinear, and difficult to explain. These characteristics prevent domain experts from understanding the reasoning behind specific decisions, a vital requirement on domains such as medical diagnosis, criminal justice, or financial decision-making.

ML algorithms' opacity brought up the need for interpretable algorithms creation, the main focus of the eXplainable Artificial Intelligence (XAI) field. According to Vilone and Longo [1], there are three motivating factors: (i) the demand to produce more transparent models; (ii) the need for techniques that enable humans to interact with them; and (iii) the requirement of trustworthiness of their inferences. Moreover, the recently approved General Data Protection Regulation (GDPR) document [2] introduced the right of explanation. These guidelines aim to give individuals the right to obtain an explanation of the inference(s) automatically produced by a model, confront, and challenge an associated recommendation, particularly when it might negatively affect an individual legally, financially, mentally, or physically.

Since explainability is an inherently human-centric property, XAI research has received increasing attention from scholars of several different domains in the research community.

As stated in [3], "currently, there is a broad array of definitions and expectations for XAI, which require multidisciplinary research efforts, as existing communities have different requirements and often have drastically different priorities and areas of specialization". For instance, ML engineers aim to create either interpretable models or explain black-box models with post-hoc techniques. Meanwhile, Human–Computer Interaction (HCI) researchers are mainly focused on building solutions that satisfy end-user needs, independently of the technical approach adopted. There are also relevant discussions on the fields of Philosophy, Psychology, and Cognitive Science, particularly in merging the existing research regarding how people generate or evaluate explanations into current XAI research [4]. This evident multidisciplinary nature makes the design and evaluation of XAI systems an intrinsically challenging task. In fact, the effectiveness of explainability is founded on the "perception and reception of the person receiving the explanation", which means that the user's explainability needs greatly influence technical choices, both in terms of design and evaluation. As a result, designing and evaluating a XAI system can be considered as much of a design challenge as an algorithmic one [5].

The contributions of this work are threefold. First, we aim to provide an overview of the most relevant evaluation methods already proposed in the literature to evaluate XAI systems. Our extensive literature review identified an urgent need to standardize the categorization and terminologies used for the different explanations' proprieties and respective evaluation methods. Thus, our second contribution focuses on providing a new taxonomy to organize the already available XAI evaluation methods clearly and intuitively. The multidisciplinary nature of XAI research was considered an essential requirement during the design of this taxonomy, which resulted in a clear separation between Human-centred and Computer-centred methods. Third, we discuss possible ways to use the proposed taxonomy in the design and evaluation of XAI systems.

The conducted literature review examined relevant papers from five prominent academic databases and bibliographic search engines, namely ScienceDirect, Engineering Village, ACM Digital Library, Arxiv and Google Scholar. To identify and select potential research articles, we used a keyword-based search using terms such as "XAI", "Explainable Artificial Intelligence", "XAI Evaluation", "Human-centred", and "Computer-centred". We restricted the research to articles published between 2017 and 2022. This initial list was then filtered in terms of relevance and quality of the papers. We then used the reference list of the selected articles to identify and include additional papers in our literature review. These papers were thoroughly examined, and an effort was made to identify the research questions the authors addressed in their works, if not clearly stated. This process supported the extraction of the XAI aspects evaluated in each revised research work, which subsequently facilitated its respective categorization according to the taxonomy presented in Section 3.

This paper is structured as follows: Section 1 summarizes the motivation and objectives of the work; Section 2 presents relevant background information for the evaluation of XAI systems; in Section 3, the proposed taxonomy for XAI systems' evaluation methods is presented; Sections 4 and 5 give an overview of the most relevant Human-centred and Computer-centred evaluation methods, respectively; Section 6 contains the takeaways from this research work and discusses how the proposed taxonomy can be used in the design and evaluation of XAI systems; and finally the conclusions and future work are drawn in Section 7.

2. Background

In this section, we address general background topics that are crucial to understand the current state of XAI evaluation methods, namely: (i) types of ML explanations; (ii) the importance of XAI evaluation; (iii) current taxonomies for XAI evaluation; and (iv) current pitfalls of XAI evaluation methods.

2.1. Types of ML Explanations

The ML explanation methods can be divided into three different types [6]:

- **Attribution-based explanations:** These type of explanations aim to rank or assign an importance value to input features based on their relevance to the final prediction. These are, arguably, some of the most common explanations to be evaluated in the literature.
- **Model-based explanations:** These explanations are represented by models used to interpret the task model. These can be the task model itself or other more interpretable post-hoc models created for that purpose. Common metrics to evaluate these are related to model size (e.g., decision trees depth or number of non-zero weights in linear models).
- **Example-based explanations:** As the name implies, Example-based explanations provide an understanding of the predictive models through representative examples or high-level concepts. When analysing specific instances, these methods can either return examples with the same prediction or with different ones (counterfactual example).

2.2. Importance of XAI Evaluation

As XAI-based solutions are becoming ever more frequent (and necessary), it is vital to properly evaluate their explainability components. This assessment can have several purposes, such as asserting that the explanations are faithful to the associated ML model, or ensuring that they are actually effective and useful to the end users. Current XAI research is mainly focused on creating new methods to improve explainability while simultaneously ensuring high predictive performance [1], often demonstrating that XAI explanations can positively impact the user understanding and Trust on an ML system. However, some authors also defend that the mere presence of explanations can cause these effects, regardless of their content [7] and consequentially give a false sense of security. Moreover, there is also an inherent human bias towards simpler explanations, which could contribute to systems with more persuasive explanatory outputs being adopted, instead of more transparent ones [8]. As such, it is imperative that explanatory methods and artifacts produced by XAI systems are thoroughly evaluated both before and throughout their deployment in a production environment.

2.3. Current Taxonomies for XAI Evaluation

Doshi-Velez and Kim [9] proposed a taxonomy based on the participation of real humans in the evaluation, and whether the task is equal to the real use case task or is a simpler version. The taxonomy is structured as follows: (a) Application-grounded evaluation (end task)—Requires conducting end user experiments within a real application; (b) Human-grounded evaluation (simple task)—Refers to conducting simpler human-subject experiments that maintain the essence of the target application. The difference is that these experiments are not carried out with the domain experts, but with laypersons; (c) Functionally-grounded evaluation (proxy task)—Requires no human experiments. In this type of evaluation, some formal definition of Interpretability serves as a proxy to evaluate the explanation quality, e.g., the depth of a decision tree. In order to link these types of evaluation, the authors describe a set of open problems and approaches to tackle them, ultimately advocating for the creation of large repositories containing problems that correspond to real-world tasks which require human input. With this structure in place, ML methods could be used to identify latent dimensions representing certain Interpretability factors.

In a more recent work, Zhou et al. [10] take this previous categorization and split the evaluation methods between the ones involving real human participation and those that do not. Complementing this, the authors go a step further by specifying the adequate metrics for the application and human-grounded evaluation methods, discerning between subjective and objective metrics. As for the functionally-grounded methods, the authors distinguish between three types of explanations (Model, Attribution or Example-based),

serving as inspiration for this work. Additionally, a set of quantitative metrics is presented together with the explainability properties they evaluate for each explanation type.

Finally, Mohseni et al. [3] published a very extensive survey on the evaluation of XAI systems. The authors believe in a multidisciplinary approach to XAI since various research fields have explored the Interpretability of ML systems and can bring significant improvements to the creation of more robust XAI techniques. Moreover, they propose a categorization for XAI design and evaluation methods consisting of two attributes: (a) design goals—which are gathered from multiple research domains and organized between three target user groups; and (b) evaluation measures—which are obtained from evaluation methods present in literature.

2.4. Shortfalls of Current XAI Evaluation

As previously mentioned, XAI evaluation is a complex problem since it relies not only on the potential technical approach adopted but also on the fulfillment of user's needs. Despite recent works with very positive outcomes towards improving the Interpretability and effectiveness of XAI systems, some pitfalls still need to be tackled.

Lack of evaluation: A recent survey by Anjomshoae et al. [11] focused on reviewing works for explainable agents and robots and indicated that 97% of the 62 evaluated articles point out that explanations serve a user need, but 41% did not evaluate their explanations with such users. Moreover, from the papers that performed a user evaluation, relatively few provided a good discussion of the context (27%), results (19%) and limitations (14%) of their experiment. Another survey from Adadi and Berrada [12] reviewed 381 papers and found that only 5% had an explicit focus on the evaluation of the XAI methods. Although the evaluation of XAI techniques is vital to ensure they fulfill the desired goals, these reviews show that only a small portion of efforts are directed towards exploring such evaluation.

Lack of consensus: When comparing the existing taxonomies for XAI evaluation, it is possible to identify a lack of consensus in several aspects, from categorization and terminology to the considered properties and evaluation metrics. In terms of categorization, Mohseni et al. [3] focus their research on Human-Centered based evaluation methods, which require the use of participant feedback. Their proposed categorization is very detailed and complete with regard to the existing evaluation methods and their respective evaluation goal. Vilone and Longo [1] and Zhou et al. [10] distinguish between Human-centred evaluation and AI-based evaluation, which relies only on the AI system itself to execute the evaluations. However, the former type of evaluations are described with a lower level of granularity that lacks the detail provided by Mohseni et al. [3]. Another example is the uneven usage of the terms "subjective" and "objective" across literature. Zhou et al. [10] and Vilone and Longo [1] state that the "subjective" term requires the involvement of a human in the explanation's evaluation, while the "objective" term concerns only evaluation aspects that do not depend on the judgements of the participants. At the same time, Mohseni et al. [3] uses them with a distinct purpose, being both used for evaluation methods that involve collecting user feedback. These conceptual incompatibilities make the usage of current taxonomies a challenging task for systematic and standardised evaluation of XAI systems.

Lack of multidisciplinarity: Another pitfall for several XAI techniques and their evaluation is the disregard of a multidisciplinary approach to the creation and evaluation of such techniques. As most methods are created on a more technical environment, they usually ignore the potential contributions that other areas like HCI might bring to the table. Similar to Mohseni et al. [3], Liao and Varshney [5] defend that a broad view is necessary for XAI research because users tend to prefer to have a holistic understanding of the system. Since explainability is an inherently human-centric property, the authors believe that the HCI can bring great contributions towards solving XAI algorithms limitations, highlighting three main arguments: Firstly, there is no one-size-fits-all solution for producing useful explanations. Therefore, the technical choice for a particular XAI method should be guided by the explainability needs of the different kinds of users, which is where HCI can offer

important insights and methodological tools; secondly, problems may arise on empirical studies with real users that technical knowledge cannot solve alone. Design approaches can be paired with the technical views in order to overcome these issues; in addition, thirdly, theories on Human Cognition and Behaviours provide conceptual tools that may motivate new computational and design frameworks for XAI.

Lack of standardized evaluation procedures: Although XAI techniques can usually be applied in several contexts, the same is not true for the evaluation procedure of such methods (in the case there is one), where there is a lack of standardized evaluation procedures that enable an efficient and exhaustive evaluation of explanations. As a result, researchers adopt new ways of evaluating explanation methods tailored to each new use case they come across, which deteriorates the ability to interpret and compare the outcomes of these experiments. Nevertheless, there has been some efforts to standardize the evaluation procedures and tools used. For instance, Quantus is an open-source toolkit proposed by Hedström et al. [13] that consists of a collection of computer-based evaluation metrics for evaluating explanation methods. Regarding human-centered evaluation procedures, the task of standardizing an evaluation procedure becomes even more challenging, due to the subjective nature and variety of the methods in this category. While some works put effort into describing the process of these evaluation methods like [1], others like [3] go a step further and propose a list of guidelines that can help navigating through a XAI evaluation procedure. Nevertheless, the effectiveness of both procedures and guidelines still lacks experimental proof in literature.

Lack of incorporation of cognitive processes: According to [11], subjective evaluation measures obtained from interviews or user feedback about Satisfaction, Trust, or explanation Usefulness are much more prevalent than objective measures. While objective measures can be less ambiguous and reliant on the user itself, subjective measures enable collecting insights on how the user perceives the explanation and the system. Nevertheless, subjective measures can mislead the design and evaluation of XAI systems if the cognitive processes of how people generate and evaluate explanations are not carefully considered. As an illustrative example, Liao and Varshney [5] recently highlighted the importance of dual-process theories [14] for XAI research, which assume that people can engage in two different systems to process information and decide upon it. While System 1 involves intuitive thinking, following mental shortcuts and heuristics created from past experiences, System 2 is analytical thinking, relying on careful reasoning of information and arguments. A wide range of current XAI techniques implicitly assume that the end-user mostly uses System 2 thinking and will attend to every bit of explanation provided. However, in reality, people are more likely to engage in System 1 thinking they lack either the ability or motivation to perform analytical thinking. Therefore, providing an extremely detailed XAI explanation does not guarantee that the end-user will use all the supplied information, which can be misleading during the evaluation of the provided explanations.

Lack of visualization and interaction strategies: In [4], the authors defend that much XAI research is based on researchers' intuitions of what constitutes a good explanation, rather than a Human-centred approach that considers user's expectations, concerns and experience. Likewise, current XAI research still does not properly address how end-users interact, visualize, and consume the information supplied by XAI systems. This lack of customized strategies significantly influences XAI evaluation procedures, since it will directly impact how end-users perceive, process, use, and consequently evaluate AI explanations.

As far as this research goes, the remarkable advances that already proposed taxonomies brought for XAI evaluation are unquestionable. Nevertheless, none of those taxonomies fully satisfied all the conditions we considered crucial on a XAI evaluation taxonomy. By building upon these previous works, in this paper, we merge, complement and standardize their different contributions, resulting in a new taxonomy for XAI evaluation that aims to foster a more systematic and standardised evaluation of XAI systems. We believe our contributions in this work can aid in the definition of more reliable and robust

XAI evaluation methods, while also increasing the community awareness regarding this critical topic. It is paramount to ensure that XAI evaluation is a priority when designing, implementing and deploying XAI systems and that it encompasses both the technical and the human interaction side of any ML system. Moreover, the proposed taxonomy can help solving the lack of standardization for terminologies and methods within the XAI topic, while illustrating the perspectives from which it is possible to conduct XAI evaluation.

3. Taxonomy for XAI Systems Evaluation Methods

A variety of terminologies and categorization for XAI evaluation methods have surged in literature as a direct consequence of the multidisciplinary nature of research efforts. Although it would be ideal to create a "one-size-fits-all" taxonomy to sort these methods, each knowledge domain looks at the XAI evaluation issue from a different perspective. As such, it is very challenging to reach a single taxonomy that encompasses knowledge from each one of those perspectives. For this reason, in this work, we decided to maintain the "multidisciplinary" essence of XAI research while adopting a new taxonomy (represented in Figure 1), whose purpose is to be a map for XAI evaluation methods during the development process of XAI solutions. The first step for building this taxonomy was to split the XAI evaluation methods into two big families: **Human-centred** and **Computer-centred** methods. While the former corresponds to methods that require conducting user experiments with human subjects, the latter involves other methods that take advantage of formal definitions of Interpretability to evaluate the quality of explanations. Furthermore, each family was divided into different categories, and each category can also have a set of sub-categories.

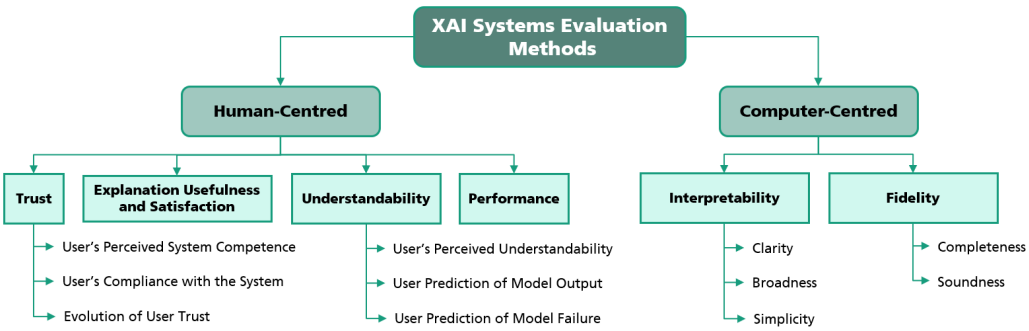

Figure 1. Proposed taxonomy for XAI systems' evaluation methods.

Human-centred evaluation methods were split into four categories, each one corresponding to the target XAI concept being evaluated (see Table 1). These concepts are specific XAI constructs considered relevant in literature from several research areas, which help paint a picture of the value added to the user experience by a XAI system. The definitions for each Human-centred category and sub-category were obtained from different reviewed works (e.g., Mohseni et al. [3], Gunning and Aha [15]) and from multidisciplinary research efforts on XAI conducted by our team.

Regarding Computer-centred methods, the categorization and respective terminologies were inspired in the work of Zhou et al. [10]. In particular, the methods were divided in two major categories, which then were split into five different sub-categories (see Table 2).

The following Sections 4 and 5 provide the state-of-the-art review for Human-centred and Computer-centred XAI evaluation methods, respectively. We structured these two sections in a similar way: (i) start by providing a brief introduction that explains the rationale behind the selection of the considered categories and sub-categories; (ii) create a sub-section for each category, where the most relevant research works for each category are briefly presented and discussed; and (iii) present a summary of all the research works reviewed (in a table format), including the most relevant works discussed in the sub-sections referred in (ii).

Table 1. Human-centred evaluation methods: categories and sub-categories.

Category	Sub-Category
Trust: a variable factor shaped by user interaction across time and usage, which affects how comfortable the user is when using the XAI system. User perception influences its beliefs on the XAI system outputs.	**User's Perceived System Competence**: depicts the user position on how capable an ML system is when solving a particular task.
	User's Compliance with the System: focused on understanding if the user would rely on the system's decision or not to act upon a task.
	Evolution of User Trust: represents how a user's Trust can vary across time and usage of a particular ML system.
Explanation Usefulness and Satisfaction: two inherently connected aspects relevant to assess user experience. The same explanation can imply different levels of Usefulness, depending on the information revealed, and can also lead to a different level of user Satisfaction.	
Understandability: the ability to outline the relation between the input and output of a particular system with respect to its parameters. It is usually defined as a user's mental model of the system and its underlying functions.	**User's Perceived Understandability**: depicts the user understanding of system's underlying functions.
	User Prediction of Model Output: focused on understanding if the user is able to define model behaviour on a particular instance or kind of data.
	User Prediction of Model Failure: focused on understanding if the user is correctly able to identify the scenarios where the system fails a particular task.
Performance: the performance of ML systems usually depends not only on the models but also on their respective users. Evaluating the performance of both agents and their interaction is essential to assess the expected performance on real scenarios.	

Table 2. Computer-centred evaluation methods: categories and sub-categories.

Category	Sub-Category
Interpretability: implies that the explanation should be understandable to humans, being important to manage the social interaction of explainability.	**Clarity**: implies that the explanation should be unambiguous.
	Broadness: describes how generally applicable is an explanation.
	Simplicity: implies that the explanation is presented in a simple and compact form.
Fidelity: implies that the explanations should accurately describe model behaviour in the entire feature space, being important to assist in verifying other model desiderata or discover new insights of explainability.	**Completeness**: implies that the explanation describes the entire dynamic of the ML model.
	Soundness: describes how correct and truthful is an explanation.

4. Human-Centred Evaluation Methods

Human-Centred Evaluation methods follow a human-in-the-loop approach, which involves the interaction of individuals with the system to test one or more of its properties. The main goal is to provide insights on the expected end-users' perspective of the system, by highlighting strengths and weaknesses of the XAI method and its application, as well as possible opportunities for improvement.

Although literature has a high diversity of techniques, there are some that are more widely used than others. Firstly, Think-Aloud is a method in which participants can freely express any thought regarding the task at hand, which might not be directly related to XAI explanations. A similar, yet more focused in explanations methods, is User Self-Explanation where participants expose their thoughts in a more structured manner, particularly towards explaining their reasoning in particular actions.

Another commonly used evaluation method is the **Interview**, where a set of pre-structured questions is asked to the participants. These questions can be either closed or open-ended and allow researchers to gather more specific information about a particular aspect of the system/method. The interviews could be conducted during or after participants interact with the system, depending on the goal. Finally, **Likert-scale questionnaires** are also widely used across literature. These questionnaires focus on evaluating specific properties with the Likert-scale, a well-known method developed by the psychologist Rensis Likert. The Likert scale is typically a five-, seven- or nine-point scale that measures the level of agreement or disagreement of the participant.

Figure 2 presents an overview of a XAI system evaluation process from a Human-centred perspective. The diagram was adapted from the work of Gunning and Aha [15]. The adaptation was necessary as the original diagram intended to represent the entire explanation process, while our work is focused solely on the evaluation of such process. Moreover, the terminology originally used did not entirely fit the structure of the taxonomy proposed in Section 3. As such, the main strategy of adaptation involved replacing the "XAI Measurement Category" boxes of the original diagram by the "Evaluation Categories" detailed in Table 1. Some relations between each Category and other components of the diagram were also reformulated.

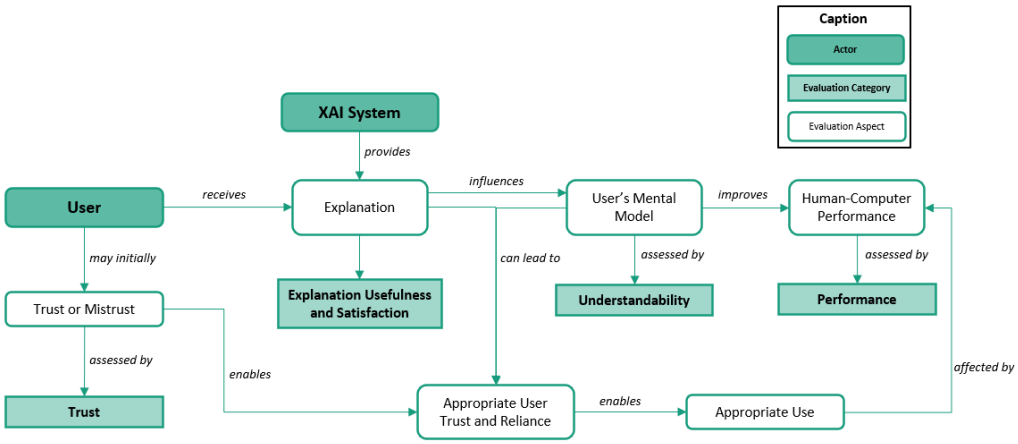

Figure 2. Human-centred evaluation process diagram (adapted from [15]) .

Sections 4.1–4.4 describe each evaluation category considered in the proposed taxonomy for the Human-centred methods, namely **Trust**, **Understandability**, **Explanation Usefulness and Satisfaction** and **Performance**, respectively. These descriptions were based on explainability concepts presented in previously works [1,3,10,15] and adapted to our taxonomy vision. The most relevant evaluation methods for each category are briefly presented and discussed, being also presented a summary of these methods and respective categorization in Section 4.5.

4.1. Trust

User Trust is a critical requirement when deploying an ML system that profoundly affects the user's perception of such system, which in turn influences the ability to believe in its outputs without fearing any danger [16]. It is inherently related to the user's confidence and to how comfortable the user is when using that system on a real task. Initially, a user's Trust is set according to prior knowledge and existing beliefs. As the user explores the system, the levels of Trust and confidence can fluctuate, depending on the use cases that the user comes across. Therefore, user Trust is a variable factor that is shaped by the user interaction, where he might experience different Trust and mistrust feelings [3].

Most scholars do not evaluate user Trust directly because it is hard to pin down what system factors influence Trust the most. Therefore, each work usually chooses a particular aspect of a XAI system and assesses its impact on user Trust.

Firstly, we can evaluate user Trust through interviews and user self-explanations during or after the user experiments with the system, like in Bussone et al. [17] and Cahour and Forzy [18]. Moreover, Likert-scales questionnaires are also used widely across literature, such as in Berkovsky et al. [19], Bussone et al. [17], Cahour and Forzy [18] and Nourani et al. [20]

Another way of assessing user Trust is through the **User's Perceived System Competence**, which shows the user position on how capable an ML system is when solving a particular task. For example, Yin et al. [21] evaluated the impact of the accuracy of an image classification model on user Trust. The results show that a user's Trust can be influenced by both the actual model's accuracy and by the model's perceived accuracy. Nourani et al. [20] also explored how the user's perception of accuracy is affected by the inclusion of explanations and their level of meaningfulness. The results indicate that it is not only important to have explanations, but also to have human-meaningful ones, since "weak" explanations that the user cannot understand might lower the user perceived accuracy. Moreover, authors also state that the "understanding of processing logic is more important for user Trust than the history of observed results alone". Therefore, the **User's Perceived Understandability** is another important factor that influences user Trust (see Understandability Section 4.2 for more details).

One could also study the **User's Compliance with the System**, which focuses on understanding if the user would rely on the system decision or not to act upon a task. Berkovsky et al. [19] evaluated the impact of several recommendation interfaces and content selection strategies on user Trust. The method considered nine key factors from Trust, divided into three categories, one of them involving explanations. Eiband et al. [7] experimented with "placebic" explanations (explanations that convey no information) in mindlessness of user behaviour. The method consisted of understanding if this type of explanation would influence user Reliance on the system. The results suggest that "placebic" explanations can increase the levels of Trust in a similar way to real explanations. The authors also propose that they could be used in a future work as "baseline" explanations that would serve as a comparison basis for other explanations.

Another way of measuring any of the above aspects of user Trust is to compare explanations generated by an AI system with ones provided by humans. Kunkel et al. [22] used this strategy on a recommender system to explore the perceived quality of explanations created by two different sources: a personal source (i.e., other humans) or an impersonal source (a machine). The results show that users tend to trust explanations created by other humans more than the explanations generated by a machine. Follow up evaluations to these results could be understanding in what way personal explanations are different from impersonal ones, and how the latter could be improved to either be more trustworthy or mimic the former.

As previously stated, Trust is not a static property. Besides the system itself, user's experience and learning over time can make the user's Trust evolve positively or negatively. Holliday et al. [23] assessed Trust and reliance in multiple stages of working with an explainable text-mining system. Results showed that, as the user became more familiar and experienced with the system, the level of Trust also changes over time. Yang et al. [24] also explored how the level of Trust evolved as the user experience with the system increased utilizing two different Trust measures on their experiments.

For this reason, monitoring the levels of users' Trust over time appears crucial when evaluating any ML system, with or without an explanatory component, to help understand how Trust progresses with experience over time.

Finally, Zhang et al. [25] conducted a case study that explores how model features can influence user Trust and the joint Performance of the Human-AI system. The study was divided into two experiments, each one with a particular goal in mind. The first experience was based on an income prediction task, where subjects had to predict whether a person's annual income would exceed $50K or not, based on a set of attributes. Three experimental factors were used across different scenarios: the AI's confidence level, the AI's prediction and the amount of user additional knowledge compared to the system. For each scenario, user Trust was measured using two behavioural indicators that are not reliant on a subjective user self-report. The first one was a "switch percentage", which shows the proportion of trials were the participant chose to use the AI's prediction instead of his own. The second on was the "agreement percentage" is the percentage of trials were both the participant's and AI's prediction was equal. Results showed that the main influence factor was whether the user had access to the AI's confidence level. In those cases, the user would switch to the AI's prediction more often. The second experience focused on the effect of local explanations on user Trust. The setup was equal to the previous experiment, but local explanations generated using SHAP values were presented to the user instead of confidence levels. Using the same indicators as in the previous experiments, results did not indicate that explanations had a significant impact on user Trust, compared to the baseline scenario.

4.2. Understandability

Understandability can be defined as "the ability to characterize the relation between the input and output of a system with respect to its parameters" [26]. For a system to be understandable, one could say that it must "support user understanding of system's underlying functions" [3].

In the cognitive psychological field, a user's understanding is usually defined as the user's mental model. Researchers in HCI analyze these users' mental models to evaluate the level of understanding of a system, a method that can be applied to ML systems. In the case of XAI systems, users come across explanations that help them build a mental model of how the system works, supposedly more accurately than in a non-XAI version of the same system. Therefore, studying these users' mental models in XAI systems is a method of assessing the effectiveness of explanations in increasing Understandability.

Explanations and their relationship with understanding have been studied for several years in human–AI interaction research, focusing mainly on figuring out what are the important characteristics of an ideal explanation [27].

Firstly, user understanding could be assessed through user expectations, which would also tell researchers what type of explanations should be generated. Rader and Gray [28] investigated user understanding of algorithmic curation in Facebook's News Feed, and whether they believed that all their friends' posts appeared on their feed. The results showed a wide range of beliefs and causal inferences which were highly dependent on the personal experience of each user. Consequently, these perception differences also affect how the user interacts with the system. Lim and Dey [29] conducted two experiments to understand what information users are interested in when interacting with a few real-world applications. The authors defined a set of types of explanations the user might want, and tried to define if that information would satisfy the user's need.

As mentioned in the User Trust section, Nourani et al. [20] stated that "understanding of processing logic is more important for user Trust than the history of observed results alone". This work assessed **User's Perceived Understandability** through an image classification review task. The experiments showed a significant difference between strong and weak explanations, where the tendency was to have lower user Trust when the subjects did not understand the explanation. Without human-meaningful explanations, people expected the system to fail more based on past observations.

Nothdurft et al. [30] also evaluated the perceived Understandability using transparency explanations and justifications on a system with "unexpected negative events". The results support the authors' hypothesis that explanations can help reduce the user's Trust loss when he faces one of those events, in comparison to an incomprehensible system that does not produce justifications for them.

User Prediction of Model Output is connected to the user expectations of a model and can be used to evaluate the level of Understandability of a model. Ribeiro et al. [31] applied a set of evaluation methods to their proposed XAI methods, LIME and SP-LIME, where one of them was focused on evaluating if the generated explanations lead users to relevant system insights. The authors intentionally trained a bad classification model, and subjects had to answer a couple questions regarding their understanding of the model output and what were the relevant features used by the model in that scenario. Results suggested that explanations were very useful in providing model insights and in deciding when not to trust the model and why. In a follow-up work [32], the same authors proposed another XAI method, Anchors, which represent local and "sufficient" conditions for a model's prediction. The user study compared these explanations with linear explanations on the same classification model and evaluated the ability of users to accurately predict the model output.

Similar to Trust, users' mental models of a system are also consolidated over time, and not on a single use. Moreover, it is common practice for developers to regularly update AI systems, whether it is training the same model with higher quality data or replacing the algorithm for a better one. Although such updates improve the system Performance on the validation set, their impact on the end-users is not linear. Bansal et al. [33] studied these updates to AI systems and how they influence both user understanding and task Performance. They introduced the concept of "compatibility" of an update, defining two score metrics to evaluate an update to a classifier, one for local and the other for global compatibility level. This compatibility was evaluated with regard to the user's mental model and how it is affected by the update.

Model Understandability could also be measured through **User Prediction of Model Failure**. Bansal et al. [34] states that with the rise of human-AI teams in high-stake domains decisions (healthcare, criminal justice, etc.), it is increasingly important to focus on the team Performance that is dependent on both the AI system and the user. The authors studied the user awareness of an AI system, particularly on the system's error boundary, which is vital when the user decided to accept or override an AI-based recommendation. The experiments' goal was to build insights on how the user model is affected by the system's error boundary and how relevant it is in the resulting team Performance. Results suggested that certain error boundary properties can influence the effectiveness and efficiency of the collaboration between an AI system and its users.

Nushi et al. [35] affirm that understanding the details of system's failure is vital for identifying areas for refinement, communicating the reliability of a system in particular scenarios, and defining appropriate human oversight and engagement. Nevertheless, the characterization of failures and shortcomings is a complex task on ML systems, and the existing evaluation methods have limited capabilities. The authors propose a set of hybrid human–machine methods named Pandora for component-based AI systems that can help list the conditions of a system's malfunction. The experimental tests on an image captioning system indicate that these methods can discover failure details that go unnoticed when using only traditional metrics.

Shen et al. [36] studied the impact of visual interpretations on understanding incorrectly predicted labels produced by image classifiers. Although one would expect that explanations would help the subjects identify incorrect labels more accurately, experimental results suggested otherwise. In fact, the group of users that had access to the explanations and the labels performed worse than the group with only the labels, which suggests that the interpretations presented were ineffective. Further investigation would be needed to figure out the reason for these results.

Finally, some studies on the literature focused on assessing Understandability through interviews and think aloud approaches [37,38], as well as Likert-scale questionnaires [39,40]. In short, Understandability is characterised by the ability to make a human understand and attribute meaning to the explanations provided by the system. While this favours increased user Trust in the system, it is not enough to make the user's task more efficient. In practice, an explanation can be understandable, but not useful or relevant to the user.

4.3. Explanation Usefulness and Satisfaction

The ultimate goal of any system is to satisfy a particular user need. Therefore, in a XAI system, it is also important to evaluate how useful explanations are to the user and how satisfied the user is with them. A large portion of literature follows a qualitative evaluation of user Satisfaction of explanations, using questionnaires and interviews [29,41,42]. One possibility to consider is to conduct expert case studies. Compared to lay people, experts have more knowledge on the subject around the system and can provide a more complete and in-depth opinion [43,44].

Some authors study the explanation Usefulness, which helps understand whether explanations are valuable to the user or not. For example, Coppers et al. [45] studied the Satisfaction level of expert translators when using either an intelligible or a non-intelligible version of the same system. The results show that the added explanations do not necessarily lead to a significant change in user experience. Participants stated that the intelligibility was only valuable when the information provided was beneficial to the translation process, or when it added information that the expert was not immediately aware of.

Besides the Usefulness of explanations, it is also important to consider the amount of information those explanations present to the user. Poursabzi-Sangdeh et al. [46] conducted a series of experiments focused on evaluating the effect of presenting different information to the user on the ability to fulfil a particular task. In a few cases, although users had access to either more information about the model or about its prediction process, they also were less able to identify and correct models' mistakes, seemingly due to information overload. Lim and Dey [29] also refer this information overload as an important problem to consider when developing systems for real-world applications, since those scenarios are usually more complex. On another work, Gedikli et al. [41] evaluated a recommender system according to some explainability goals. One of those was efficiency, which can be defined as how good an explanation is in reducing the decision time. The authors concluded that explanations helped users decide more quickly, but they do not guarantee there would be no implications on the decision quality or the user Satisfaction.

Despite the interpretability benefits of explanations, they also bring costs to the user. In the case that the explanations are too extensive, the user might need more time to process the information presented. Bunt et al. [47] investigated this aspect by conducting two studies on low-cost decision support systems: one on the comprehensibility and perceived cost of explanations and the other on the user desire for explanations. The results show that most users are interested in having a "sufficient" transparency level, while access to too much information leads to a viewing cost that outweighs the benefit of the explanations.

Lim et al. [42] studied the effectiveness of different types of explanations. On the one hand, the users had "why" explanations, which showed why the system behaved in a certain way. On the other hand, the users had access to "why not" explanations, which showed why the system did not behave in a certain way. Results showed that the former contributes to a better understanding and stronger feeling of trust on the model when compared to the latter. This strategy could be used when researchers want to compare case-based explanations versus counterfactual explanations, which correspond to the "why" and "why not" scenarios.

Thus, Understandability and Usefulness are mutually complementary explanation's attributes. The latter enables us to assess an explanation's real benefit towards users' needs. Ultimately, the Usefulness of explanations may translate into gains in users' Task Performance, another metric we now describe.

4.4. Performance

One of the XAI research main goals is to support end-users' decisions and help them be more successful in the task they have at hand. Therefore, Task Performance is an important measure to consider. Although Performance can be evaluated separately on the model and end-user, it is much more relevant to observe the overall Performance of the Human-AI system, since it is in that condition that systems will function in real-life scenarios. There is not a lot of quantity and diversity in Performance evaluation of XAI systems. Most human-in-the-loop approaches are focused on the Explainability concepts described in the previous sections. Nevertheless, there are still ways of including Performance in an evaluation process.

A more direct approach to Performance evaluation would be to compare the Performance in two scenarios: one where the user had access to the system's output generated explanations and other where only the system's output was known. Despite having a different goal in mind, this approach has been described in previous sections. It would only be necessary to take the experiment outcomes and look at them from a Performance perspective. For example, consider a case with two subject groups where one had access to a classification model output and explanation and the other only was able to see the output. If each subject had to decide to either agree or disagree with the model's decision, the accuracy of each subject could be calculated and compared between the two subject groups.

Bansal et al. [33] studied these updates to AI systems and how they influence both user understanding and task Performance. A more detailed description of this work was provided in a previous Section 4.2. Through the proposed metrics, authors were able to evaluate the trade-off between the Performance and the compatibility of an AI system update.

While evaluating any Human-AI system, researchers can also take advantage of evaluation results to improve the current Model Performance. For instance, when adopting a method focused on the User Prediction of Model Failure, one could take the chance to identify the cases where the model fails and retrain it on more data like those same cases. Moreover, this could also be useful to identify cases where the generated explanations fail and opens the opportunity for improving them in future iterations. Ribeiro et al. [31] concluded that users were able to detect wrong explanations in a text classification model, which enabled the training of better classifiers, with improved Performance and explanations quality.

4.5. Summary

Table 3 summarizes the previously described state-of-the-art methods by presenting a condensed view of the research questions, evaluation methods/metrics used and the type of XAI explanation generated for each work. Each row represents a Method/Category pair, as there are situations where the same work proposes different metrics for different categories.

Table 3. Human-centred evaluation methods for machine learning explanations.

Reference	Category	Sub-Category	Research Question	Methods and/or Metrics	XAI Type
Coppers et al., 2018 [45]	Explanation Usefulness and Satisfaction	-	What is the impact of intelligibility on the Perceived Value, Trust, Usage, Performance, and User Satisfaction for different translation aids?	**SUS questionnaire** and **5-point Likert scale** to evaluate Perceived Usefulness, general Usability and the perception of the following features: understandable, useful, enjoyable, trustworthy, improves quality and efficiency.	Attribution-based
Bunt et al., 2012 [47]	Explanation Usefulness and Satisfaction	-	To what extent does the participant want to know more about how the system generates its intelligent behaviour?	**Qualitative interviews**, **2-week diary study** and **7-point Likert scale** to evaluate perceived utility, perceived accuracy and matching expectations.	Example-based
Lim et al., 2009 [42]	Explanation Usefulness and Satisfaction"	-	How to evaluate if different types of explanations improve the users perception of system's Usefulness and Satisfaction?	**Survey** that asked users to explain how the system works and to report their perceptions of the explanations and system in terms of Usefulness, Satisfaction, Understandability, and Trust.	Model-based
Ribeiro et al., 2016 [31]	Performance	-	Can non-experts improve a classifier through explanations?	Iterative model Performance assessment through explanations based on feature importance.	Attribution-based
Ribeiro et al., 2018 [32]	Performance	-	How do explanations influence user Performance when trying to predict model behaviour on unseen instances?	**Coverage:** fraction of instances predicted after seeing the explanation; **Precision:** Fraction of correct predictions; **Time:** Seconds the user took to complete the task per prediction.	Attribution-based
Bansal et al., 2019 [34]	Performance	-	How updates to an AI system can affect human-AI team Performance?	**ROC:** Team Performance metric; **Compatibility score:** fraction of examples on which the older model version recommends the correct action, the new model version also recommends the correct action; **Locally-compatible update:** Whether the action given by the new model version affect the a user's mental model created during an older model version usage; **Globally-compatible update:** Update is compatible for all mental-models (users).	Model-based
Lim et al., 2009 [42]	Performance	-	How to evaluate if different types of explanations lead to better task Performance?	**Task Performance**: evaluated total learning time and average time completion.	Model-based
Holliday et al., 2016 [23]	Trust	Evolution of User Trust	How does user trust evolve over time?	**7-point Likert scale** to indicate the extent of agreement with a statement of trust in the system before and after performing a task assisted by the system; **Think aloud protocol**: through which authors identified four factors of trust: perceived system ability, perceived control, perceived predictability, and perceived transparency.	Attribution-based

Table 3. *Cont.*

Reference	Category	Sub-Category	Research Question	Methods and/or Metrics	XAI Type
Yang et al., 2017 [24]	Trust	Evolution of User Trust	How does user trust evolve and stabilize over time as humans gain more experience interacting with automation?	**TrustEND**: trust rating elicited after the terminal trial T); **TrustAUTC**: Area Under the Trust Curve; and **Trust*e***: Trust of entirety; **Response Rates (RR)** and **Response Times (RT)** of system Reliance (trusting the automation in the absence of threat alarms) and system compliance (trusting the automation in the presence of one or more threats).	Model-based
Eiband et al., 2019 [7]	Trust	User's Compliance with the System	Do placebic explanations invoke similar levels of Trust as real explanations?	**5-point Likert scale** questionnaire to evaluate user's perception of Trust.	Attribution-based
Nourani et al., 2019 [20]	Trust	User's Perceived System Competence	How local explanations influence user perception of model's accuracy?	**Implicit perceived accuracy**: percentage of responses where participants predicted correct system classifications; **Explicit perceived accuracy**: users' numerical estimate of the system's accuracy on a scale 0 to 100%.	Attribution-based
Kunkel et al.., 2019 [22]	Trust	User's Perceived System Competence	How to evaluate the impact of explanations on user Trust in recommender systems?	**5-point rating scale** to evaluate explanations quality.	Example-based
Yin et al.., 2019 [21]	Trust	User's Perceived System Competence	How model's stated accuracy affects Trust?	**Agreement fraction**: percentage of tasks in which users' final prediction matched model's predictions; **Switch fraction**: percentage of tasks in which the users revised their predictions to match model's predictions.	Model-based
Zhang et al., 2020 [25]	Trust	User's Perceived System Competence	How Trust, Accuracy and Confidence score on Trust calibration are affected by: (1) showing AI's prediction versus not showing, and (2) knowing to have more domain knowledge than the AI?	**Switch percentage**: how often participants chose the AI's predictions as their final predictions); **Agreement percentage**: trials in which the participant's final prediction agreed with the AI's prediction).	Model-based
Samuel et al., 2021 [48]	Trust, Understandability	User Prediction of Model Output	What is the impact of showing AI Performance and predictions' explanations on people's Trust in the AI and the decision outcome?	Human subject experiments with **5-Likert scale surveys** to evaluate predictability, reliability and consistency.	Attribution-based
Kim et al., 2018 [39]	Trust, Understandability	User Prediction of Model Output	How to quantitatively evaluate what information saliency maps are able to communicate to humans?	**10-point Likert scale** to evaluate how important they thought the image and the caption were to the model. **5-point Likert scale** for evaluating how confident they were in their answers. Evaluated accessibility, customization, plug-in readiness and global quantification.	Model-based

Table 3. *Cont.*

Reference	Category	Sub-Category	Research Question	Methods and/or Metrics	XAI Type
Shen et al., 2020 [36]	Understandability	User Prediction of Model Failure	How useful is showing machine-generated visual interpretations in helping users understand automated system errors?	**Accuracy** of human inferences on model misclassification.	Attribution-based
Nushi et al., 2018 [35]	Understandability	User Prediction of Model Failure	How detailed Performance views can be beneficial for analysis and debugging?	**Human Satisfaction:** Indicates whether the user agrees with the image / caption pair presented; **System Performance / Prediction Accuracy:** Accuracy of the system prediction compared to ground truth (defined by human Satisfaction).	Model-based
Ribeiro et al., 2016 [31]	Understandability	User Prediction of Model Output	Do explanations lead to insights?	A **counter** of how many models each human subject trusts, and **open-ended questions** to indicate their reasoning behind their decision.	Attribution-based
Ribeiro et al., 2018 [32]	Understandability	User Prediction of Model Output	How do explanations influence user Understandability when trying to predict model behaviour on unseen instances?	**Coverage:** fraction of instances predicted after seeing the explanation; **Precision:** Fraction of correct predictions; **Time:** Seconds the user took to complete the task per prediction.	Attribution-based
Nourani et al., 2019 [20]	Understandability	User's Perceived Understandability	How human perceived meaningfulness of explanation affects their perception of model accuracy?	**Post-study questionnaire** and **think-aloud approach** to evaluate implicit perceived accuracy and explicit perceived accuracy.	Attribution-based
Nothdurft et al., 2014 [30]	Understandability	User's Perceived Understandability	How different explanations goals affect human-computer Trust?	Evaluate **Perceived Understandability**, **Perceived Technical Competence**, **Perceived Reliability**, **Personal Attachment** and **Faith**.	Attribution-based
Lim et al., 2009 [42]	Understandability	User's Perceived Understandability	How to evaluate if different types of explanations help users better understand the system?	**User understanding**: evaluated correctness and detail of reasons participants provided by participants in a Fill-in-the-Blanks test and a Reasoning test.	Model-based

5. Computer-Centred Evaluation Methods

Human-centred methods are more commonly applied to evaluate XAI systems than Computer-centred, partially due to the complexity associated with judging certain properties such as Trust or Understandability from a non-human point of view. Nevertheless, Human-centred approaches also have identifiable drawbacks. In particular, Herman [8] indicates that the human bias towards simpler explanations can result in more persuasive explanations as opposed to transparent systems [3]. As such, explanation evaluation methods disconnected from the human user must also be considered. Doshi-Velez and Kim [9] categorize these methods as functionally grounded evaluations, suggesting that they are better employed after human validation of the model or system, with the exceptions of when these are not yet mature enough or human experimentation is unethical.

Two commonly found properties used to evaluate this type of explanations are their **Interpretability** and **Fidelity**. Interpretability can be described as the ability to explain in understandable terms to a human [9], while Fidelity indicates how accurately a model's behaviour is described by an explanation [6]. These properties can be further subdivided into more specific characteristics. More specifically, within the Interpretability category, it is also possible to identify the **Broadness** sub-property, which measures how generally applicable an explanation is [10] as well as the **Simplicity** (also known as Parsimony) and **Clarity** characteristics, with the former indicating if an explanation is presented in a compact or not too complex form and the latter the unambiguity associated with an explanation [6]. As for the Fidelity property, we can sub-divide it into the **Soundness** and **Completeness** categories. The first informs on how truthful an explanation is to the task model, while the last indicates if an explanation provides sufficient information to compute the output for a given input [6]. According to [1], Completeness is a measure of how many input features that affect the decision process are captured in an explanation.

These properties are, at times, also accompanied by other ones aimed at evaluating the XAI system as well, such as the Sensitivity to input perturbation and model parameter randomization. ElShawi et al. [49] did precisely this, evaluating the LIME, anchors, SHAP, LORE, ILIME and MAPLE Interpretability frameworks, albeit with different methods and metrics. More specifically, they based their work on the three axioms proposed by Honegger [50]. These pertain to the **Identity** (identical instances must have identical explanations), **Stability** (instances belonging to the same class must have comparable explanations) and **Separability** (dissimilar instances must have dissimilar explanations) of the XAI system. It should be noted that these proprieties could be considered ramifications of the Soundness sub-category. Thus, we considered that including a third hierarchical level in our proposed taxonomy would bring an unnecessary level of complexity that could compromise its practical use.

Given this, we considered that the best approach to structure the Computer-centred evaluation methods would be to follow the categorization already proposed by Zhou et al. [10]. Sections 5.1 and 5.2 describe each evaluation category considered in the proposed taxonomy for the Computer-centred methods, namely Interpretability and Fidelity, respectively. These descriptions were based on explainability concepts presented in previous works [6,9,10] and adapted to our taxonomy vision. The most relevant evaluation methods for each category are briefly presented and discussed, a summary of these methods and respective categorization also presented in Section 5.3.

5.1. Interpretability

The non-human assessment of the Interpretability of a XAI system can be more challenging than its Fidelity, in part due to the inherent subjectivity of what is considered an interpretable explanation. Nevertheless, several authors propose distinct methods to obtain a measure of this property. Nguyen and Martínez [51], for example, define the Effective Complexity metric which, when having a low value, is an indication of simple and broad explanations. They also indicate a Diversity measure which reflects the degree of integration of an explanation. Other common metrics to evaluate this property are related

to the model size, such as the Depth of decision trees or the number of non-zero weights in linear models, as noted by Ribeiro, Singh and Guestrin [31]. Slack et al. [52] used a simple metric as well to assess the Interpretability of their models: the number of runtime or arithmetic operations for a given input.

Alternatively, Hara and Hayashi [53] utilize the number of regions in which the input space is divided as a different metric to judge the complexity of their simpler models used as interpretable versions of tree ensembles. Deng [54] also proposed a framework for interpreting tree ensembles, using several metrics to evaluate the rules extracted, namely the Frequency of a rule (or the proportion of instances satisfying the rule condition), its Error (the quotient between the number of incorrectly and correctly classified instances by the rule) and its Complexity or length, defined as the number of variable-value pairs in its condition.

Lakkaraju et al. [55] proposed a framework as well, although aimed at interpreting black box models through learned decision sets (if-then rules). The evaluation of these rules/explanations involves simplistic Interpretability metrics, such as the total number of rules in the set, the maximum width of all the elements in it, the total number of predicates and unique neighbourhood descriptors, and the number of instances which satisfy two different rules, for every pair of rules. In an earlier but similar work (also based on decision sets), Lakkaraju et al. [40] evaluated the Interpretability of the framework through other metrics, such as the degree of overlapping between every pair of rules in the set, the fraction of data points not covered by any rule, the average rule length (number of constituting predicates), the total number of rules and the fraction of class labels that are predicted by at least one rule. In another work, Bhatt, Weller and Moura [56] also define several metrics for evaluating XAI explanations, one of which aims at assessing their complexity by computing the entropy of the fractional contribution of every feature to the total magnitude of the attribution. In another work, Bau et al.'s [57] proposed a framework aimed at a different scenario, more specifically the evaluation of CNN latent representations (activation maps). This is accomplished by measuring the intersection between the internal convolutional units and pixel-level semantic concepts previously annotated in the image. In this manner, if there is a high overlap between a unit and a concept in several images, it can be said that the unit represents that particular concept.

Zhang, Wu and Zhu [58] used an evaluation metric previously proposed in [59] for similar cases denoted as location instability. In their work, the feature map of a filter f is first computed and the inference location of f determined by the convolutional unit with the highest activation score (whose receptive field's centre is backpropagated to the image plane). From here, the in-image distance between this location and certain ground-truth landmarks is registered. For example, in order to investigate if a filter represents a cat, some landmarks could be its head, legs and tail, as the distance from these to a perceived activation centre should remain stable. In this manner, it is possible to judge whether a filter consistently represents the same concept by analysing the deviation between each activation-landmark pair calculated across several images.

5.2. Fidelity

The assessment of XAI system Fidelity seems to be significantly more addressed in the literature when compared with its Interpretability, probably due to its more objective nature.

For instance, Nguyen and Martínez [51] present several pertinent metrics for this task as well, such as the monotonicity associated with the feature attributions, calculated through the Spearman's correlation coefficient between these values and the corresponding estimated expectations (computed using the outputs of the model). In this manner, it is possible to assess if the attributed feature importance values are correct by checking if they are monotonic or not. They also evaluate the Broadness of these explanations by using the mutual information between a random variable and its corresponding value after feature extraction. The lower this value, the broader the explanation. In addition to this, the authors also propose a non-sensitivity measure to ensure that only features to which

the model is not functionally dependent on can be assigned a zero-importance value. They present a non-representativeness metric as well which measures the Fidelity associated with an explanation, although high values can also indicate factual inaccuracy.

Laugel et al. [60] proposed a method for measuring the risk of generating unjustified counter-factual examples, i.e., examples that do not depend on previous knowledge and are instead artefacts of the classifier, through the number of such unjustified and justified examples encountered in the neighbourhood of a particular example. In a different approach, Ribeiro, Singh and Guestrin [31] proposed a type of explanations referred to as local interpretable model-agnostic explanations (LIME) and multiple ways to evaluate them. These explanations can be seen as intrinsically explainable models (and the properties possible to retrieve from them) fitted on slight variations of the instance being analysed. They evaluate if these approximations are faithful to the model by generating a gold set of features relevant to it (which was itself interpretable in that case) and computing how many were retrieved in the explanations. Afterwards, the trustworthiness of the individual predictions and model were also assessed by deeming a set of features as untrustworthy in the first case and adding noisy features in the second, followed by the analysis of the predictions of the model in the face of these changes.

Taking inspiration from the previous work, Plumb et al. [61] introduced a method towards regularizing models for explanation quality at training time. To accomplish this, it made use of two separate metrics, namely the Fidelity of the explanations, characterized by how well an explainable model approximates the original one around a particular instance, and their Stability, which can be described by the degree to which the explanation changes between the instance's generated neighbourhood points. These metrics are approximated through two algorithms and linked with a standard predictive loss function, encouraging the model to be interpretable around a particular instance.

In the already mentioned work of Bhatt, Weller and Moura [56], another metric is proposed by the authors to evaluate the faithfulness of an explanation by iteratively replacing random subsets of given attributions with baseline values and then measuring the correlation between the differences in function outputs and the sum of these attribution subsets. Similarly, Alvarez-Melis and Jaakkola [62] also measure this property by calculating the correlation between the model's Performance drops when removing certain features from its input and the relevance scores (attributions) on various points. In another work [63], the authors define an additional method to evaluate the robustness of the explanatory model through a local Lipschitz estimate.

Contrastingly, Sundararajan, Taly and Yan [64] propose a set of axioms the XAI method(s) must satisfy instead of measuring individual properties. The first one is related to the Sensitivity of the explanations, stipulating that "(...) for every input and baseline that differ in one feature but have different predictions than the differing feature should be given a non-zero attribution". Moreover, they also add a complement to this axiom indicating that if a model's function does not mathematically depend on some variable, then the attribution associated with that variable is always zero. The second axiom is that of implementation invariance, meaning that the attributions generated by a method should always be identical for functionally equivalent models or networks (systems whose outputs are equal for the same inputs). This axiom is similar to Montavon, Samek and Müller's [65] definition of continuity, stating that nearly equivalent points should also have nearly equivalent explanations, something which can be assessed by measuring the maximum variation of the explanation in the input domain. Kindermans et al. [66] also defined an axiom related to the invariance of the explanations but in this case in relation to the model's input, indicating that an explanation method (in this case saliency-based) needs to mirror the Sensitivity of the predictive model with respect to transformations of the input.

Kim et al. [39] introduced concept activation vectors (CAVs) for quantitatively measuring the Sensitivity of black-box model predictions to previously defined concepts at a given layer of the model. According to the authors, a CAV is essentially the normal to a

hyperplane separating examples with and without a concept in the model's activations and can be represented by a binary linear classifier. They expand on this by introducing another measure representing the fraction of a class inputs whose layer activation concept vector was positively influenced by a certain concept, comparing the obtained results with their own saliency map evaluations for validation.

Ylikoski and Kuorikoski [67] also touch on the subject of **Sensitivity**, defining it as a measure of how much an explanation changes in the face of background alterations. More precisely, they consider that, as Sensitivity increases, the explanatory relationship weakens. Similarly, Yeh et al. [68] consider a variant of this property known as max-sensitivity which measures the the maximum change in the explanation with a small perturbation of the input. The authors also present an explanation infidelity measure based on the notion of the goodness of an explanation, or rather its ability to capture how the predictor function changes in response to significant perturbations. Deng et al. [69] used this metric to evaluate the attribution methods tested, as well as the ratio of the number of bounding-box pixels that had a high value attributed to them, representing the localization accuracy. Similarly, Kohlbrenner et al. [70] evaluate attribution localization in object detection problems by computing the ratio between the sum of positive relevance inside a bounding box and the total positive sum of relevance in the image. The authors also defined a weighted variant of this metric to avoid numerical issues in edge cases related to the bounding box size.

Hooker et al. [71] proposed a different approach to measure the accuracy of attribution estimates in deep neural networks aimed at image classification tasks. Their method, referred to as "Remove and Retrain" or ROAR, involves replacing the top t most relevant pixels (according to the importance estimates) with the per-channel mean, effectively removing them from the original image, and through this process generating new training and test datasets on which the model is re-trained and evaluated. From there, the accuracy of the new model is calculated and compared with the original one: if there was a noticeable decrease, then it is likely that the removed inputs were informative and that the importance estimates were accurate; if not, then the removed inputs were either uninformative or redundant, and therefore the importance estimates were not of high quality. A similar method was also proposed by Samek et al. [72] in which the most relevant regions/pixels were replaced with values randomly sampled from a normal distribution.

Adebayo et al. [73] evaluate explanations based on saliency maps through two techniques. The first consists of comparing the output of one of the saliency methods on a trained model with the output of the saliency method on a randomly initialized untrained network with the same architecture. Similarly, in the second technique, the authors measure the distance between the outputs of a saliency method trained on a given dataset and the outputs of the same method trained on the same architecture but with a copy of the dataset containing randomly permutated labels. Finally, Ignatiev [74] proposed a distinct explanation system based on if-then rules, or rigorous explanations, obtained through abductive reasoning. The validity of the outputted explanations was assessed by computing the percentage of incorrect and redundant rules.

5.3. Summary

The previously described state of the art methods are summarized in Table 4, as well as the corresponding research questions, evaluation methods/metrics and XAI explanation type outputted. Each row represents a Method/Category pair, as there are situations where the same work proposes different metrics for different categories.

Table 4. Computer-centred evaluation methods for machine learning explanations.

Reference	Category	Sub-Category	Research Question	Methods and/or Metrics	XAI Type
Lakkaraju et al., 2016 [40]	Fidelity	Completeness	How to evaluate Completeness of rule-based models?	**Fraction of classes**: Measures what fraction of the class labels in the data are predicted by at least one rule (optimal value is 1 - every class is described by some rule).	Model-based
Ignatiev, 2020 [74]	Fidelity	Completeness	How to evaluate if explanations hold in the entire instance space?	**Correctness**: percentage of correct, incorrect and redundant explanations.	Model-based
Nguyen et al., 2007 [51]	Fidelity	Soundness	How to measure the strength and direction of association between attributes and explanations?	**Monotonicity**: Spearman's correlation between feature's absolute Performance measure of interest and corresponding expectations.	Attribution-based
Nguyen et al., 2007 [51]	Fidelity	Soundness	How data processing changed the information content of the original samples (target-level analysis)?	**Target Mutual Information**: Measured between extracted features and corresponding target values (e.g., class labels).	Attribution-based
Nguyen et al., 2007 [51] Ylikoski et al., 2010 [67]	Fidelity	Soundness	How robust is an explanation to unimportant details?	**Non-sensitivity**: Cardinality of the symmetric difference between features with assigned zero attribution and features to which the model is not functionally dependent on.	Attribution-based
Bhatt et al., 2020 [56]	Fidelity	Soundness	Does the explanation captures which features the predictor used to generate an output?	**Faithfulness**: Correlation between the differences in function outputs and the sum of random attribution subsets replaced with baseline values.	Attribution-based
Bhatt et al., 2020 [56]	Fidelity	Soundness	How sensitive are explanation functions to perturbations in the model inputs?	**Sensitivity**: If inputs are near each other and their model outputs are similar, then their explanations should be close to each other.	Attribution-based
Alvarez-Melis et al., 2018 [62]	Fidelity	Soundness	Are relevance features scores indicative of "true" importance?	**Faithfulness**: Correlation between the model's Performance drops when removing certain features and the attributions.	Attribution-based
Alvarez-Melis et al., 2018 [63] Plumb et al., 2019 [61]	Fidelity	Soundness	How consistent are the explanations for similar/neighboring examples?	**Robustness**: Local Lipschitz estimate.	Attribution-based
Sundararajan et al., 2017 [64]	Fidelity	Soundness	How sensitive are explanation functions to small perturbations in the model inputs?	**Sensitivity**: Measures the degree to which the explanation is affected by insignificant perturbations from the test point.	Attribution-based
Sundararajan et al.,2017 [64]	Fidelity	Soundness	Are attributions identical for functionally equivalent networks with different implementations?	**Implementation invariance**: Measures the similarity between explanations provided by two functionally equivalent networks.	Attribution-based
Montavon et al., 2018 [65]	Fidelity	Soundness	How fast the prediction value goes down when removing features with the highest relevance scores?	**Selectivity**: Measures the ability of an explanation to give relevance to variables that have the strongest impact on the prediction value.	Attribution-based

Table 4. *Cont.*

Reference	Category	Sub-Category	Research Question	Methods and/or Metrics	XAI Type
Kindermans et al., 2019 [66]	Fidelity	Soundness	Is the explanation method input invariant, i.e., mirrors the behaviour of the predictive model with respect to transformations of the input?	**Input invariance**: Method that demonstrates that there is at least one input transformation that causes a target explanation method to attribute incorrectly.	Attribution-based
Yeh et al., 2019 [68]	Fidelity	Soundness	How sensitive are explanation functions to small perturbations in the model inputs?	**Max-sensitivity**: Calculated based on the maximum change in the explanation when adding small perturbations to the input.	Attribution-based
Yeh et al., 2019 [68]	Fidelity	Soundness	Does the explanation method captures how the predictor function changes in the face of perturbations?	**Explanation infidelity**: Expected difference between the dot product of the input perturbation to the explanation and the output perturbation (i.e., the difference in function values after significant perturbations on the input).	Attribution-based
Kohlbrenner et al., 2020 [70]	Fidelity	Soundness	Does the explanation method reflect the object understanding of the model closely, i.e., both predictions and explanations are only based on the object itself?	**Attribution localization**: Ratio between the sum of positive relevance inside a bounding box and the total positive sum of relevance in the image.	Attribution-based
Hooker et al., 2018 [71]	Fidelity	Soundness	How to avoid that distribution shift influences the estimated feature importance?	**Remove and retrain (ROAR)**: Remove the data points estimated to be most important, and retraining the model to measure the degradation of model Performance.	Attribution-based
Samek et al., 2016 [72]	Fidelity	Soundness	How predictions change when the most relevant data points are progressively removed?	**Region perturbation via MoRF (Most Relevant First)**: Measure how the class encoded in the image disappears when the information is progressively removed from the image using an ordered sequence of locations by relevance.	Attribution-based
Adebayo et al., 2018 [73]	Fidelity	Soundness	How to assess similarity between two visual explanations?	**Spearman rank correlation** with absolute value (absolute value), and without absolute value (diverging); **SSIM**: The structural similarity index; **Pearson correlation**: Correlation of the histogram of gradients (HOGs) derived from two maps.	Attribution-based
Kim et al., 2018 [39]	Fidelity	Soundness	How to quantify the concept importance of a particular class?	**TCAV score (TCAVq)**: measures the positive and negative influence of a defined concept on a particular activation layer of a model.	Attribution-based
Laugel et al., 2019 [60]	Fidelity	Soundness	How to evaluate the behaviour of a post-hoc Interpretability method in the presence of counterfactual explanation?	**Justification score**: Binary score that equals 1 if the counterfactual explanation is justified, 0 if unjustified; **Average justification score**: Average value of the justification score computed over multiple instances and multiple runs.	Example-based
Ribeiro et al., 2016 [31]	Fidelity	Soundness	How to evaluate if explanations retrieve the most important features for the model?	**Recall of important features**: Measures the amount of a gold set of features considered important by the model that are recovered by the explanations.	Model-based

Table 4. *Cont.*

Reference	Category	Sub-Category	Research Question	Methods and/or Metrics	XAI Type
Ribeiro et al., 2016 [31]	Fidelity	Soundness	How explanations can be used to select between competing models with similar Performance?	**Trustworthiness**: Prediction analysis after adding noisy and untrustworthy features.	Model-based
Lakkaraju et al., 2017 [55]	Fidelity	Soundness	How to evaluate if transparent approximations used as explanations capture the black-box model behaviour in all parts of the feature space?	**Disagreement**: Percentage of predictions in which the label assigned by the model explanation does not match the label assigned by the black box model.	Model-based
Plumb et al., 2019 [61]	Fidelity	Soundness	How each feature influences the model's prediction in a certain neighborhood?	**Neighborhood-fidelity (NF)**: Accuracy of the model in a certain neighborhood.	Model-based
Nguyen et al., 2007 [51]	Fidelity	Completeness	How to measure the representativeness of Example-based explanations?	**Non-representativeness**: Performance measure of interest (e.g.,cross-entropy) between the predictions of interest and model outputs, divided by the number of examples.	Example-based
Nguyen et al., 2007 [51]	Interpretability	Broadness and Simplicity	How data processing changed the information content of the original samples (feature-level analysis)?	**Feature mutual information**: Measured between original samples and corresponding features extracted for explanations.	Attribution-based
Nguyen et al., 2007 [51]	Interpretability	Broadness and Simplicity	How to assess the effects of non-important features?	**Effective complexity**: Minimum number of attribution-ordered features that can meet an expected Performance measure of interest.	Attribution-based
Montavon et al., 2018 [65]	Interpretability	Clarity	If model response to certain data points are nearly equivalent, are the respective explanations also nearly equivalent?	**Continuity**: Measures the strongest variation of the explanation in the input domain.	Attribution-based
Bau et al., 2017 [57]	Interpretability	Clarity	How to evaluate explanations' Clarity using human-labeled visual concepts?	**Network Dissection**: Measure the intersection between the internal convolutional units and pixel-level semantic concepts previously annotated in an image.	Attribution-based
Zhang et al., 2018 [58]	Interpretability	Clarity	How to evaluate explanations' Clarity through the semantic meaningfulness of CNN filters?	**Location instability**: Measured by the distance between known image landmarks and inference locations (areas with highest activation scores).	Attribution-based
Lakkaraju et al., 2017 [55]	Interpretability	Clarity	How to measure the unambiguity of transparent approximations used to explain rule-based models?	**Rule overlap**: For every pair of rules, sum up all the instances which satisfy both rules' conditions simultaneously (optimal value is zero); **Cover**: The number of instances which satisfy the condition of the target rule (optimal value is the size of the entire dataset).	Model-based

Table 4. *Cont.*

Reference	Category	Sub-Category	Research Question	Methods and/or Metrics	XAI Type
Lakkaraju et al., 2016 [40]	Interpretability	Clarity	How to evaluate Clarity of rule-based models?	**Fraction uncovered**: Computes the fraction of data points which are not covered by any rule.	Model-based
Bhatt et al., 2020 [56]	Interpretability	Simplicity	How complex is the explanation?	**Complexity**: Entropy of the fractional contribution of every feature to the total magnitude of the attribution.	Attribution-based
Nguyen et al., 2007 [51]	Interpretability	Simplicity	How to measure the diversity of Example-based explanations?	**Diversity**: Distance function in the input space between different examples, divided by the number of examples.	Example-based
Ribeiro et al., 2016 [31]	Interpretability	Simplicity	How to measure the complexity of explanations?	**Complexity**: Decision tree depth, number of non-zero weights in linear models.	Model-based
Hara et al., 2016 [53]	Interpretability	Simplicity	How to measure the Interpretability of tree ensembles?	**Tree ensembles complexity**: Measured by the number of regions in which the input space is divided.	Model-based
Deng, 2019 [54]	Interpretability	Simplicity	How to measure the quality of rules extracted from tree ensembles?	**Rule complexity**: Measured by the length of the rule condition, defined as the number of variable-value pairs in the condition; **Rule frequency**: Proportion of data instances satisfying the rule condition; **Rule error**: Number of incorrectly classified instances determined by the rule divided by the number of instances satisfying the rule condition.	Model-based
Lakkaraju et al, 2017. [55]	Interpretability	Simplicity	How to measure the complexity and intuitive representation of transparent approximations used to explain rule-based models?	**Size**: The number of model rules; **Max width**: The maximum width computed over all the elements; **No. of predicates**: Number of predicates (appearing in both the decision logic rules and neighborhood descriptors); **No. of descriptors**: the number of unique neighborhood descriptors; **Feature overlap**: For every pair of a unique neighborhood descriptor and decision logic rule, number of features that occur in both.	Model-based
Lakkaraju et al., 2016 [40]	Interpretability	Simplicity	How to evaluate Simplicity of rule-based models?	**Average rule length**: The average number of predicates a human reader must parse to understand a rule.	Model-based
Slack et al., 2019 [52]	Interpretability	Simplicity	How to evaluate simulatability, i.e., user's ability to run an explainable model on a given input?	**Runtime operation counts**: Measure the number of Boolean and arithmetic operations needed to run the explainable model for a given input.	Model-based

6. Discussion

Nowadays, XAI tools and techniques are more accessible than ever, for researchers and practitioners to build XAI solutions for several applications. The demand for explanations expanded beyond researchers that aim to comprehend the models better, to end-users of the developed XAI system to increased Trust and AI adoption. Nevertheless, there is still work to be done regarding how users perceive explanations generated by machines, and what their impact is on real world scenarios. Therefore, it is crucial to ensure reliable and trustworthy practices are applied to assess the impact of such explanations on the end-users' decision-making. However, the evaluation of explanations has been neglected in a significant number of works found in literature. On the one hand, there are works that do not conduct a complete validation of their proposed methods, nor do they compare them to other state of the art methods. On the other hand, when works describe the evaluation process for the proposed explanation methods, it is tailored to the specific context and target user, making it difficult to apply them in other scenarios.

This review gathers works that contributed with methods and/or insights of the evaluation of explanation methods. The proposed taxonomy is a structure for organizing such methods, which can be extended to other XAI properties, encompassing both subjective and objective categories with Human-centred and Computer-centred categories, respectively. The ability to organize the spectrum of evaluation methods into a single taxonomy brings several advantages to future research works. Firstly, a taxonomy condenses disperse knowledge into a categorization scheme that generally makes concepts more accessible. Secondly, it solves the already identified issue of having to understand different terminologies from different authors, since it standardizes the concepts. Finally, it also helps researchers identifying potential research gaps in this field, thus motivating the community to tackle new challenges that were unknown before.

Due to the lack of standardized evaluation procedures and the diversity of methods and metrics, identifying the most commonly used ones is a very extensive and challenging process. Even though it was possible to join distinct works into each category, the majority of them adopt methods and metrics purposely created for each work, for both Human-Centred and Computer-Centred Evaluation methods. The only exceptions are Likert-scale questionnaires from the Human-Centred family, which is used in 8 out of the 23 works highlighted in the respective section.

The following takeaways resulted from an analysis on the reviewed methods, which provide valuable guidelines for any evaluation process of XAI systems. For instance, in the context of software development, it is very common to iterate over systems deployed in a production environment, either to add new features or correct older ones, as per the necessities of the end user. This is even more vital in ML systems, since they are dependent on real world data that tend to change overtime, a phenomenon known as data drift. An ML system update can significantly impact model behaviour, and consequently the way users interact with the entire ML system. For high-stake systems, it is imperative to ensure that the impact of such changes does not negatively affect not only the system itself, but its users' interaction. For example, tunning an explanation method for a specific class might demand from the user more time/effort to learn/trust the new model's behaviour, even if it results in a system's performance improvement from a technical perspective.

Even without model updates, humans (users of ML systems) are in a constant learning process. As a consequence, users' mental models and properties like **Trust** depend on the user's system knowledge and familiarity. Moreover, real-world scenarios are also everchanging, which might expose the ML system to new and unseen data, leading to unexpected outcomes. Therefore, continuous monitoring after deployment of XAI solutions is very important to ensure that the quality and validity of explanations evolve as expected across time and system usage.

For evaluation methods in general, it is common to set a baseline scenario, used as a point of comparison to others with different conditions. The goal is to understand if these different conditions lead to a different outcome from the baseline case. The choice

of the baseline scenario can greatly impact the meaning of the results. The majority of approaches reviewed adopt a "no explanation" baseline, where the participants do not have access to the model explanations [31,45]. In another case, one of the experiments described in Poursabzi-Sangdeh et al. [46] asked users to complete a particular task before even consulting the model. The usage of different baselines adds another level of complexity to the discussion of experimental results from the same or different works, as there is the need analyse and compare them in a fair setting. As the reviewed works do not provide a clear view of what experimental settings are vital for a reliable and comparable comparison between them, we believe it should be a high priority for future works to clearly define the evaluation settings considered, while taking into account already available knowledge from other works within the same subject. This would enable more transparent and comparable research works for both publications' authors and readers.

Despite the controlled nature of evaluation experiments, there might still exist certain human behavioural differences between scenarios that are not reflected on the evaluation results. In a study by Langer et al. [75] about people's behaviour when presented with explanations, the results showed that the inclusion of explanations increased the participants' willingness to comply with a request, even if those explanations did not contain relevant information. Eiband et al. [7] denominated these explanations as "placebic explanations" and argued that placebic and real explanations could lead to a similar level of user trust in the system. The authors also stated that works on explainability techniques might benefit from using placebic explanations as a baseline scenario, since it would eliminate any influence the presence of explanations might have on the user. Another study by Buçinca et al. [76] evaluated two commonly used techniques for evaluating XAI systems. One of the experiments conducted tested the usage of "proxy tasks", a simplified version of a real task for a particular system. The results support the hypothesis that the outcomes from proxy tasks may not be a good indicator for the performance of the XAI system in a realistic setting. Therefore, it is important to ensure that the metrics and measures used to evaluate experimental results are as representative as possible of the several aspects of the XAI system undergoing evaluation.

One could argue that the limitations of existing evaluation methods could be a consequence of the nature of XAI techniques themselves. Liao and Varshney [5] raised concerns regarding this same topic. Although efforts have been yield to approximate ML systems to their users through explainability, there is still a clear disconnection between technical XAI approaches and their real effectiveness in supporting users objectives for their particular use case. Although it is not considered on the proposed taxonomy, the kind of end-users and their respective needs from a XAI system is an important aspect that should be taken into account during the development and evaluation process. Users can differ not only on their role when interacting with the system, but also in their domain knowledge, expertise with ML systems or even in which environment will they use the XAI system. Furthermore, research also shows that issues arise when the assumptions underlying XAI technical approaches are detached from people's cognitive processes. One of the possibilities is that the sole presence of explanations can result in unwarranted trust in models, as it was demonstrated by Langer et al. [75] and Eiband et al. [7]. For these reasons, the evaluation of such approaches should account for these conditions; otherwise, its results might be misleading.

From the three main types of explanations previously mentioned that can be evaluated through these methods, Attribution-based explanations are arguably the easiest to assess since the metrics used to perform their evaluation are usually independent from the methods or models used to obtain the importance values. Contrastingly, methods to evaluate Model-based explanations can be highly dependent on the target XAI approach due to its intrinsically specific characteristics. There are, although, exceptions to this when different model types can obtain similar explanations, such as if-then rule constructs. As for Example-based explanations, their evaluation shares some similarities with Attribution-

based explanations in the sense that, in theory, the same evaluation method could be applied to different systems, although not as straightforward as with the latter.

Limitations

This paper consists of a general review of articles on XAI evaluation methods. The aim was to find articles that provided either Human-centred or Computer-centred methods. We found it relevant to find forms of evaluation exclusively linked to the performance and comprehensibility of a XAI system. Thus, we did not specifically look for studies assessing the whole system Relevance and Usefulness. Rather, we looked for studies assessing the Usefulness of explanations towards increasing human's understanding and trust. For this reason, industries and/or researchers not applying a Human-centred design approach to the system conceptualization and development require addressing a complete evaluation of a system. This paper does not provide methods for that.

We also did not look for methods for assessing how individual beliefs and user's prior knowledge would affect their experience with a XAI system. The proposed taxonomy lists assessment methods for characteristics intrinsic to the XAI system itself, whether assessed by computational or human-centred methods. Thus, we only included characteristics that affect the user's ability to understand, trust and feel satisfied with a XAI system.

Focusing on the user dimension, this taxonomy does not assign methods to different user profiles. It is likely that an ML engineer will have different explanation needs compared to a domain-expert user. While different methods may better serve different profiles of users, we did not establish these relationships as we had no context to test it.

With regard to the operationalisation of the methods surveyed, this paper provides an accessible basis to facilitate the choice of methods, but does not provide a flow to support the selection of the most suitable methods for a specific XAI-based use case. Additionally, this work does not provide support on how to apply the identified methods, since this would imply replicating the content of the cited articles. To go deeper on how to apply each method, we suggest the reader to consult the methods section of the respective references.

7. Conclusions and Future Work

In this work, we reviewed several studies on XAI evaluation, which provide relevant methods and insights on the topic. We proposed a taxonomy that structures the surveyed evaluation methods into multiple categories, each one representing a valuable property of XAI systems. As this is currently an expanding field of research, it is natural that future works present new and unseen outcomes that this taxonomy does not include, but that can nevertheless be added as an extension for the current state of this work, in the form of new methods or even a new evaluation category. For instance, emotional analysis could be a suitable alternative to the qualitative methods highlighted in the Human-Centred family, such as Interviews or questionnaires. Kaklaushkas et al. [77] point out certain theories which argue that an individual's actions can be influenced by several behavioural factors, such as physical, social, psychological or emotional. Therefore, the emotional state of users might affect their interaction with XAI systems. Thus, XAI evaluation procedures might be improved using emotional analytics methods, by providing insights on how a participant's emotional state affected its levels of Trust, Understandability or other properties of XAI systems.

Moreover, this taxonomy mainly addresses two of the several shortfalls identified in Section 2.4. Firstly, it helps solve the lack of consensus between different works when defining and categorizing explanation evaluation methods. Some of the concepts and ideas behind the proposed frameworks have different meanings in different works, which create a challenging task when trying to get a broader picture of the topic. Secondly, it follows the multidisciplinary nature of XAI, which requires knowledge required from different fields. As the remaining shortfalls are considered out of the scope of this work, it is important that future works on this topic address them in a comprehensive manner.

Furthermore, in Section 6, we highlight a set of relevant aspects for XAI evaluation that are scarcely addressed in the literature. Those discussion points can be a starting point for future works to address, fostering knowledge transfer between researchers to speed up and improve research on the topic.

Still in terms of future work, we find it particularly relevant to test and map potential dependencies between different methods. For instance, try to identify if it would be relevant to apply a specific method before carrying out another, e.g., apply a set of Human-centred methods to help in the selection of the most relevant Computer-centred proprieties to be evaluated or vice versa. Additionally, such tests may contribute to define recommendations on how to interpret the results obtained for each category and sub-category of the proposed taxonomy, in a way that enables defining the necessary actions for improving the XAI system under analysis. As a final note, we also believe that this work can serve as a basis to further develop a set of guidelines and decision flows that support XAI system developers in the selection of the most suitable evaluation methods for their specific use case.

Author Contributions: Conceptualization, P.L., E.S. and L.R.; methodology, P.L., E.S., C.B. and L.R.; investigation, P.L., E.S., C.B. and L.R.; formal analysis, P.L., E.S., C.B. and L.R.; data curation, P.L., E.S., C.B. and L.R.; writing—original draft preparation, P.L. and E.S.; writing—review and editing, P.L., E.S., C.B. and L.R.; visualization, P.L., E.S., C.B. and L.R.; supervision, L.R.; project administration, L.R. and T.O.; funding acquisition, L.R. and T.O. All authors have read and agreed to the published version of the manuscript.

Funding: This work was financially supported by the project Transparent Artificial Medical Intelligence (TAMI), co-funded by Portugal 2020 framed under the Operational Programme for Competitiveness and Internationalization (COMPETE 2020), Fundação para a Ciência and Technology (FCT), Carnegie Mellon University, and European Regional Development Fund under Grant 45905.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

XAI	eXplainable Artificial Intelligence
ML	Machine Learning
HCI	Human–computer interaction

References

1. Vilone, G.; Longo, L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf. Fusion* **2021**, *76*, 89–106. [CrossRef]
2. General Data Protection Regulation (GDPR)–Official Legal Text. Available online: https://gdpr-info.eu/ (accessed on 15 May 2022).
3. Mohseni, S.; Zarei, N.; Ragan, E.D. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Trans. Interact. Intell. Syst. (TiiS)* **2021**, *11*, 1–45. [CrossRef]
4. Miller, T. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artif. Intell.* **2019**, *267*, 1–38. [CrossRef]
5. Liao, Q.V.; Varshney, K.R. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv* **2022**, arXiv:2110.10790.
6. Markus, A.F.; Kors, J.A.; Rijnbeek, P.R. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *J. Biomed. Inform.* **2021**, *113*, 103655. [CrossRef] [PubMed]
7. Eiband, M.; Buschek, D.; Kremer, A.; Hussmann, H. The Impact of Placebic Explanations on Trust in Intelligent Systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*; ACM: Glasgow, UK, 2019; pp. 1–6. [CrossRef]
8. Herman, B. The promise and peril of human evaluation for model interpretability. *arXiv* **2017**, arXiv:1711.07414.
9. Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv* **2017**, arXiv:1702.08608.

10. Zhou, J.; Gandomi, A.H.; Chen, F.; Holzinger, A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics* **2021**, *10*, 593. [CrossRef]
11. van der Waa, J.; Nieuwburg, E.; Cremers, A.; Neerincx, M. Evaluating XAI: A Comparison of Rule-Based and Example-Based Explanations. *Artif. Intell.* **2021**, *291*, 103404. [CrossRef]
12. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
13. Hedstr˙om, A.; Weber, L.; Bareeva, D.; Motzkus, F.; Samek, W.; Lapuschkin, S.; H´ohne, M.M.C. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations. *arXiv* **2022**, arXiv:2202.06861.
14. Kahneman, D. *Thinking, Fast and Slow*; Macmillan: New York, NY, USA, 2011.
15. Gunning, D.; Aha, D. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Mag.* **2019**, *40*, 44–58. [CrossRef]
16. Bhatt, U.; Xiang, A.; Sharma, S.; Weller, A.; Taly, A.; Jia, Y.; Ghosh, J.; Puri, R.; Moura, J.M.F.; Eckersley, P. Explainable Machine Learning in Deployment. *arXiv* **2020**, arXiv:1909.06342.
17. Bussone, A.; Stumpf, S.; O'Sullivan, D. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In Proceedings of the 2015 International Conference on Healthcare Informatics, Dallas,TX, USA, 21–23 October 2015; pp. 160–169. [CrossRef]
18. Cahour, B.; Forzy, J.F. Does Projection into Use Improve Trust and Exploration? An Example with a Cruise Control System. *Saf. Sci.* **2009**, *47*, 1260–1270. [CrossRef]
19. Berkovsky, S.; Taib, R.; Conway, D. How to Recommend? User Trust Factors in Movie Recommender Systems. In Proceedings of the 22nd International Conference on Intelligent User Interfaces, Limassol, Cyprus, 13–16 March 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 287–300. [CrossRef]
20. Nourani, M.; Kabir, S.; Mohseni, S.; Ragan, E.D. The Effects of Meaningful and Meaningless Explanations on Trust and Perceived System Accuracy in Intelligent Systems. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Washington, DC, USA, 28–30 October 2019; Volume 7, pp. 97–105.
21. Yin, M.; Wortman Vaughan, J.; Wallach, H. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 1–12. [CrossRef]
22. Kunkel, J.; Donkers, T.; Michael, L.; Barbu, C.M.; Ziegler, J. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; ACM: Glasgow, UK, 2019; pp. 1–12. [CrossRef]
23. Holliday, D.; Wilson, S.; Stumpf, S. User Trust in Intelligent Systems: A Journey Over Time. In Proceedings of the 21st International Conference on Intelligent User Interfaces, Sonoma, CA, USA, 7–10 March 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 164–168. [CrossRef]
24. Yang, X.J.; Unhelkar, V.V.; Li, K.; Shah, J.A. Evaluating Effects of User Experience and System Transparency on Trust in Automation. In Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, Vienna, Austria, 6–9 March 2017; ACM: Vienna Austria, 2017; pp. 408–416. [CrossRef]
25. Zhang, Y.; Liao, Q.V.; Bellamy, R.K.E. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-assisted Decision Making. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; ACM: Barcelona, Spain, 2020; pp. 295–305. [CrossRef]
26. Gleicher, M. A Framework for Considering Comprehensibility in Modeling. *Big Data* **2016**, *4*, 75–88. [CrossRef]
27. Madsen, M.; Gregor, S. Measuring Human-Computer Trust. In Proceedings of the 11th Australasian Conference on Information Systems, Brisbane, Australia, 6–8 December 2000; pp. 6–8.
28. Rader, E.; Gray, R. Understanding User Beliefs About Algorithmic Curation in the Facebook News Feed. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Seoul, Korea, 18–23 April 2015; ACM: Seoul, Korea, 2015; pp. 173–182. [CrossRef]
29. Lim, B.Y.; Dey, A.K. Assessing Demand for Intelligibility in Context-Aware Applications. In Proceedings of the 11th International Conference on Ubiquitous Computing, Orlando, FL, USA, 30 September–3 October 2009; Association for Computing Machinery: New York, NY, USA, 2009; pp. 195–204. [CrossRef]
30. Nothdurft, F.; Richter, F.; Minker, W. Probabilistic Human-Computer Trust Handling. In Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Philadelphia, PA, USA, 18–20 June 2014; Association for Computational Linguistics: Philadelphia, PA, USA, 2014; pp. 51–59. [CrossRef]
31. Ribeiro, M.T.; Singh, S.; Guestrin, C. " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
32. Ribeiro, M.T.; Singh, S.; Guestrin, C. Anchors: High Precision Model-Agnostic Explanations. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32. [CrossRef]
33. Bansal, G.; Nushi, B.; Kamar, E.; Weld, D.S.; Lasecki, W.S.; Horvitz, E. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33; pp. 2429–2437. [CrossRef]

34. Bansal, G.; Nushi, B.; Kamar, E.; Lasecki, W.; Weld, D.S.; Horvitz, E. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. In Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing, Washington, DC, USA, 28 October 2019; Volume 7; p. 10.
35. Nushi, B.; Kamar, E.; Horvitz, E. Towards Accountable AI: Hybrid Human-Machine Analyses for Characterizing System Failure. *arXiv* **2018**, arXiv:1809.07424.
36. Shen, H.; Huang, T.H.K. How Useful Are the Machine-Generated Interpretations to General Users? A Human Evaluation on Guessing the Incorrectly Predicted Labels. *arXiv* **2020**, arXiv:2008.11721.
37. Kulesza, T.; Stumpf, S.; Burnett, M.; Wong, W.K.; Riche, Y.; Moore, T.; Oberst, I.; Shinsel, A.; McIntosh, K. Explanatory Debugging: Supporting End-User Debugging of Machine-Learned Programs. In Proceedings of the 2010 IEEE Symposium on Visual Languages and Human-Centric Computing, Leganes, Spain, 21–25 September 2010; pp. 41–48. [CrossRef]
38. Binns, R.; Van Kleek, M.; Veale, M.; Lyngs, U.; Zhao, J.; Shadbolt, N. 'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; pp. 1–14. [CrossRef]
39. Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; sayres, R. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; Volume 80; pp. 2668–2677.
40. Lakkaraju, H.; Bach, S.H.; Leskovec, J. Interpretable decision sets: A joint framework for description and prediction. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1675–1684.
41. Gedikli, F.; Jannach, D.; Ge, M. How Should I Explain? A Comparison of Different Explanation Types for Recommender Systems. *Int. J.-Hum.-Comput. Stud.* **2014**, *72*, 367–382. [CrossRef]
42. Lim, B.Y.; Dey, A.K.; Avrahami, D. Why and Why Not Explanations Improve the Intelligibility of Context-Aware Intelligent Systems. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Boston, MA, USA, 4–9 April 2009; Association for Computing Machinery: New York, NY, USA, 2009; pp. 2119–2128. [CrossRef]
43. Kahng, M.; Andrews, P.Y.; Kalro, A.; Chau, D.H. ActiVis: Visual Exploration of Industry-Scale Deep Neural Network Models. *arXiv* **2017**, arXiv:1704.01942.
44. Strobelt, H.; Gehrmann, S.; Pfister, H.; Rush, A.M. LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks. *arXiv* **2017**, arXiv:1606.07461.
45. Coppers, S.; Van den Bergh, J.; Luyten, K.; Coninx, K.; van der Lek-Ciudin, I.; Vanallemeersch, T.; Vandeghinste, V. Intellingo: An Intelligible Translation Environment. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, Montreal, QC, Canada, 21–26 April 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 1–13. [CrossRef]
46. Poursabzi-Sangdeh, F.; Goldstein, D.G.; Hofman, J.M.; Wortman Vaughan, J.W.; Wallach, H. Manipulating and Measuring Model Interpretability. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 1–52.
47. Bunt, A.; Lount, M.; Lauzon, C. Are Explanations Always Important?: A Study of Deployed, Low-Cost Intelligent Interactive Systems. In Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces, Lisbon, Portugal, 14–17 February 2012; Association for Computing Machinery: New York, NY, USA, 2012; pp 169–178.
48. Samuel, S.Z.S.; Kamakshi, V.; Lodhi, N.; Krishnan, N.C. Evaluation of Saliency-based Explainability Method. *arXiv* **2021**, arXiv:2106.12773.
49. ElShawi, R.; Sherif, Y.; Al-Mallah, M.; Sakr, S. Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Comput. Intell.* **2021**, *37*, 1633–1650. [CrossRef]
50. Honegger, M. Shedding light on black box machine learning algorithms: Development of an axiomatic framework to assess the quality of methods that explain individual predictions. *arXiv* **2018**, arXiv:1808.05054.
51. Nguyen, A.; Martínez, M. On Quantitative Aspects of Model Interpretability. *arXiv* **2020**, arXiv:2007.07584.
52. Slack, D.; Friedler, S.A.; Scheidegger, C.; Roy, C.D. Assessing the local interpretability of machine learning models. *arXiv* **2019**, arXiv:1902.03501.
53. Hara, S.; Hayashi, K. Making tree ensembles interpretable. *arXiv* **2016**, arXiv:1606.05390.
54. Deng, H. Interpreting tree ensembles with intrees. *Int. J. Data Sci. Anal.* **2019**, *7*, 277–287. [CrossRef]
55. Lakkaraju, H.; Kamar, E.; Caruana, R.; Leskovec, J. Interpretable & explorable approximations of black box models. *arXiv* **2017**, arXiv:1707.01154.
56. Bhatt, U.; Weller, A.; Moura, J.M. Evaluating and aggregating feature-based model explanations. *arXiv* **2020**, arXiv:2005.00631.
57. Bau, D.; Zhou, B.; Khosla, A.; Oliva, A.; Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6541–6549.
58. Zhang, Q.; Wu, Y.N.; Zhu, S.C. Interpretable convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8827–8836.
59. Zhang, Q.; Cao, R.; Shi, F.; Wu, Y.N.; Zhu, S.C. Interpreting cnn knowledge via an explanatory graph. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.

60. Laugel, T.; Lesot, M.J.; Marsala, C.; Renard, X.; Detyniecki, M. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. *arXiv* **2019**, arXiv:1907.09294.
61. Plumb, G.; Al-Shedivat, M.; Cabrera, A.A.; Perer, A.; Xing, E.; Talwalkar, A. Regularizing black-box models for improved interpretability. *arXiv* **2019**, arXiv:1902.06787.
62. Alvarez Melis, D.; Jaakkola, T. Towards robust interpretability with self-explaining neural networks. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 1087–1098.
63. Alvarez-Melis, D.; Jaakkola, T.S. On the robustness of interpretability methods. *arXiv* **2018**, arXiv:1806.08049.
64. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017; Volume 70; pp. 3319–3328.
65. Montavon, G.; Samek, W.; M´uller, K.R. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.* **2018**, *73*, 1–15. [CrossRef]
66. Kindermans, P.J.; Hooker, S.; Adebayo, J.; Alber, M.; Sch¨utt, K.T.; D´ahne, S.; Erhan, D.; Kim, B. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 267–280.
67. Ylikoski, P.; Kuorikoski, J. Dissecting explanatory power. *Philos. Stud.* **2010**, *148*, 201–219. [CrossRef]
68. Yeh, C.K.; Hsieh, C.Y.; Suggala, A.; Inouye, D.I.; Ravikumar, P.K. On the (in) fidelity and sensitivity of explanations. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 10967–10978.
69. Deng, H.; Zou, N.; Du, M.; Chen, W.; Feng, G.; Hu, X. A Unified Taylor Framework for Revisiting Attribution Methods. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, DC, USA, 2–9 February 2021; Volume 35, pp. 11462–11469.
70. Kohlbrenner, M.; Bauer, A.; Nakajima, S.; Binder, A.; Samek, W.; Lapuschkin, S. Towards best practice in explaining neural network decisions with LRP. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–7.
71. Hooker, S.; Erhan, D.; Kindermans, P.J.; Kim, B. A benchmark for interpretability methods in deep neural networks. *arXiv* **2018**, arXiv:1806.10758.
72. Samek, W.; Binder, A.; Montavon, G.; Lapuschkin, S.; M´uller, K.R. Evaluating the visualization of what a deep neural network has learned. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *28*, 2660–2673. [CrossRef]
73. Adebayo, J.; Gilmer, J.; Muelly, M.C.; Goodfellow, I.; Hardt, M.; Kim, B. Sanity Checks for Saliency Maps. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 9505–9515.
74. Ignatiev, A. Towards Trustable Explainable AI. In Proceedings of the 29th International Joint Conference on Artificial Intelligence, Yokohama, Japan, 7–15 January 2021; pp. 5154–5158.
75. Langer, E.; Blank, A.; Chanowitz, B. The Mindlessness of Ostensibly Thoughtful Action: The Role of "Placebic" Information in Interpersonal Interaction. *J. Personal. Soc. Psychol.* **1978**, *36*, 635–642. [CrossRef]
76. Buçinca, Z.; Lin, P.; Gajos, K.Z.; Glassman, E.L. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In Proceedings of the 25th International Conference on Intelligent User Interfaces, Cagliari, Italy, 17–20 March 2020; pp. 454–464. [CrossRef]
77. Kaklauskas, A.; Jokubauskas, D.; Cerkauskas, J.; Dzemyda, G.; Ubarte, I.; Skirmantas, D.; Podviezko, A.; Simkute, I. Affective analytics of demonstration sites. *Eng. Appl. Artif. Intell.* **2019**, *81*, 346–372. [CrossRef]