data

MDPI

Review

The Role of Human Knowledge in Explainable AI

Andrea Tocchetti * and Marco Brambilla

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, 20133 Milano, Italy; marco.brambilla@polimi.it
* Correspondence: andrea.tocchetti@polimi.it

Abstract: As the performance and complexity of machine learning models have grown significantly over the last years, there has been an increasing need to develop methodologies to describe their behaviour. Such a need has mainly arisen due to the widespread use of black-box models, i.e., high-performing models whose internal logic is challenging to describe and understand. Therefore, the machine learning and AI field is facing a new challenge: making models more explainable through appropriate techniques. The final goal of an explainability method is to faithfully describe the behaviour of a (black-box) model to users who can get a better understanding of its logic, thus increasing the trust and acceptance of the system. Unfortunately, state-of-the-art explainability approaches may not be enough to guarantee the full understandability of explanations from a human perspective. For this reason, human-in-the-loop methods have been widely employed to enhance and/or evaluate explanations of machine learning models. These approaches focus on collecting human knowledge that AI systems can then employ or involving humans to achieve their objectives (e.g., evaluating or improving the system). This article aims to present a literature overview on collecting and employing human knowledge to improve and evaluate the understandability of machine learning models through human-in-the-loop approaches. Furthermore, a discussion on the challenges, state-of-the-art, and future trends in explainability is also provided.

Keywords: explainable AI; human-in-the-loop; human knowledge; explainability; traceability; interpretation; understandability; machine learning; blackbox algorithms

Citation: Tocchetti, A.; Brambilla, M. The Role of Human Knowledge in Explainable AI. *Data* **2022**, *7*, 93. https://doi.org/10.3390/data7070093

Academic Editors: Giuseppe Ciaburro and Joaquín Torres-Sospedra

Received: 18 May 2022
Accepted: 29 June 2022
Published: 6 July 2022

1. Introduction

The widespread use of Machine Learning (ML) models demonstrated its effectiveness in supporting humans in various contexts like medicine, economics, computer science and many more, while driving a never-seen technological advancement. The efficiency of such systems on both general and domain-specific tasks has driven the development of models capable of achieving even higher performance. For example, the recent development of Deep Learning and Deep Neural Networks (DNN) outperformed state-of-the-art models accuracy- and performance-wise on various tasks, such as image classification, text translation, etc. Despite the widespread excitement carried by such accomplishments, the scientific community quickly understood that ML systems could not rely on performance alone. Indeed, most complex, high-performing machine learning models were missing an essential feature. Due to their intricacy, their behaviour was not understandable to the users employing them, consequently leading to a loss of trust in such systems. Models lacking such a trait are usually referred to as black-box models, i.e., models with either known or observable input and output and hard-to-understand behaviour. These models are opposed to white-box models, i.e., systems with known or readily understandable behaviours. Such a fundamental distinction brought forth the necessity of developing methodologies to faithfully represent the logic applied by (black-box) models in a human-understandable fashion. The Explainable AI (XAI) research field poses this objective as its primary focus. Given the profound differences in how humans and machine learning systems learn, explain and represent knowledge, bridging the gap between model and human behaviour

is another fundamental objective of interest. Therefore, not only is it essential to faithfully describe model behaviour, but also to properly shape it to make it understandable to humans. For this reason, human-in-the-loop approaches have been widely employed, directly engaging the crowd to collect (structured) knowledge to evaluate and improve the interpretability of models and their explanations. Moreover, achieving a faithful, complete and understandable representation of the behaviour of a machine learning system would not only increase human trust and acceptance. Indeed, it would also be helpful to debug such systems, allowing researchers to understand their faults and consequently driving models' performance even higher. While increasing users' trust in models can be achieved by explaining their behaviour, other sources of uncertainty may influence humans' confidence in machine learning systems. Model uncertainty either comes from the inability of the model to suitably explain the data (epistemic uncertainty), from the presence of noise in the observations (aleatoric uncertainty), or from the predicted output (predictive uncertainty). A variety of approaches have been developed to solve and quantify model uncertainty [1], consequently contributing to increase model trustworthiness and detect scenarios in which explanations and model inspection are needed. In addition to increasing human trust, there are many reasons to explain the behaviour of machine learning models, like justifying its decisions, increasing its transparency to identify potential faults, improve the model, or extract new knowledge, relationships, and patterns [2]. In recent times, there has been a focus on explainability aimed at making explicit causal relationships between a model's inputs and its predictions. Such an objective is especially relevant when these relationships are not evident to the end-users employing the system or hard to understand. Moreover, such explanations provide users with a causal understanding [3] of the reasons certain input features contribute to a prediction.

Despite the call for explainability, there are still ongoing discussions on whether and when explainability is needed. Concerning such an interesting topic, Holm [4] states that the usage of black-box models is motivated *when they produce the best result, when the cost of a wrong answer is low, or when they inspire new ideas*. Another scenario in which explainability is not mandatory is low-stakes scenarios where trusting a model without understanding its behaviour would not cause any harm, even if it would misbehave. Even in high-stakes scenarios, there are some conditions and situations in which explaining the behaviour of the system is not fundamental. It is particularly true in the medical field. *If an AI model yields accurate predictions that help clinicians better treat their patients, then it may be useful even without a detailed explanation of how or why it works* ("Should AI Models Be Explainable? That depends"—https://hai.stanford.edu/news/should-ai-models-be-explainable-depends, accessed on 2 June 2022). Moreover, experiments [5,6] have revealed that providing explanations about a model's behaviour may end up generating unmotivated trust in the model. Consequently, it is fundamental to understand the role of explainability depending on the context in which the model to explain or inspect is applied and the scope in which the model deserves trust even without explainability [7].

This article provides an overview of the state of the art on the role and the contribution of human knowledge in the context of explainability of machine learning models and explainable AI. In particular, we cover methods collecting and employing knowledge to create, improve, and evaluate the explainability of blackbox models in AI. We frame such a context from the human perspective, focusing on methodologies whose main objective is to employ human knowledge as part of an explainability process. The rest of this article is structured as follows. Section 2 describes the fundamental definitions provided in the explainability research field while discussing and contextualizing their features. An overview of different methods and approaches found in the state of the art of explainability and explainable AI is also summarised. Section 3 illustrates the process applied to collect and filter the articles considered in this review. Section 4 presents the various explainability-related tasks in which human knowledge and involvement played a fundamental role, describing their approaches and discussing the findings from the literature. Section 5 summarises the article's content and describes open challenges in explainable machine learning.

2. Explainability and Explainable AI

Why can we not blindly trust a high accuracy model? Why do humans need to understand ML models? Answering these questions is not as straightforward as it may seem. There are several reasons that motivate the need to explain the behaviour of machine learning systems [2], e.g., understanding a model's logic would allow its developers to improve its performance; bank employees would be able to justify the reasons behind the rejection of a loan when such a decision is based on a model's prediction, etc. From a broader perspective, the main reason it is crucial to accurately understand the behaviour of such systems is that the unjustified application of their predictions might negatively impact our lives. In her book "Weapons of Math Destruction" [8], Cathy O'Neil describes and analyses real-life scenarios in which the improper usage of AI and machine learning models—mainly due to unjustified trust in the model—negatively affected people's lives. In particular, she emphasises that opacity is one of the three features characterising the so-called "Weapons of Math Destruction". Such a statement implicitly suggests that the application of machine learning models lacking transparency or instruments to explain their behaviour may lead to severe consequences.

2.1. Definitions

The central concept associated with Explainable AI is the notion of "explanation". An explanation can be defined as an "*interface between humans and a decision-maker that is, at the same time, both an accurate proxy of the decision-maker and comprehensible to humans*" [9]. Such a description highlights two fundamental features an explanation should have. It must be accurate, i.e., it must faithfully represent the model's behaviour, and comprehensible, i.e., any human should be able to understand the meaning it conveys. Such properties highlight the two sides of explainability: humans and models. Models should be trained to exhibit their behaviour (directly or through explainability techniques) while maintaining high accuracy and performance. A human interpreter should be capable of understanding the explanation provided by the model or explainability method. In summary, the objective of an explanation is to bridge the gap between these two worlds.

Such a dualism between human understanding and model explainability can be observed in various definitions available in the literature. In their summary of XAI, Arrieta et al. [10] provide the following characterisation of Explainable Artificial Intelligence.

> "*Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand.*"

Such a characterisation makes a series of fundamental assertions. First of all, it clearly states that the algorithm must be able to "produce details or reasons to make its functioning clear or easy to understand". This statement exemplifies the so-called self-explaining systems, i.e., models producing their output and corresponding explanations simultaneously (e.g., decision trees and rule-based models). Such systems are either inherently explainable or trained using both data and its explanations (i.e., human rationale) [11], generating models able to explain their behaviour. In the second place, Arrieta et al. consider the "audience" as a relevant entity, thus acknowledging that the interpreter influences the understandability of an explanation. Indeed, understanding how to shape explanations properly [12] is as essential as understanding how they are perceived by the audience [12–14]. For example, while an AI expert would probably prefer a detailed description of the model, a non-expert user would likely favour a small set of examples [15] representing the system's behaviour. The last aspect addressed in the definition is that the explanation must be "clear or easy to understand". Unfortunately, the concept of "easy to understand" is not the same for everyone. Indeed, it may depend on various human-related factors, such as the user's expertise with AI and ML systems, the context in which they are born, and many more [16]. Therefore, it is fundamental to properly understand how to tailor explanations depending on the audience's characteristics. Moreover, such a definition depicts a system inherently able to explain its behaviour. It does not explicitly consider models requiring the application of so-called post-hoc explainability techniques, i.e., methods able to explain the

behaviour of a ML system after its outcome has been computed. This distinction is just the first of many dimensions used to classify explainability approaches. Other categorisations classify models depending on (i) whether they are able to explain the whole model's prediction process (global explainability) or a single prediction instance (local explainability); (ii) whether they can be applied to all types of models (model agnostic) or a specific type only (model specific); (iii) the shape of the explanations (e.g., decision rules, saliency maps, etc.) and many more [9].

Before overviewing the most recent achievements in Explainable AI, it is essential to shed light on the various facets of explainability and explainable systems. Such a clarification is necessary since different research studies frame the problem from different but similar perspectives addressing distinct aspects related to explainability [2]. Among these perspectives, we claim interpretability and understandability are the most important ones described in the literature as they are strictly associated with the human side of explainability. Interpretability is defined as *"the ability to explain or provide meaning in understandable terms to a human"* [9]. Arrieta et al. [10] provide a similar definition for comprehensibility. Understandability is *"the characteristic of a model to make a human understand its function (i.e., how the model works) without any need for explaining its internal structure or the algorithmic means by which the model processes data internally"* [10]. Despite the plethora of definitions of explainability-related concepts in the literature, the final aim of the XAI research field can be summarised as developing inherently explainable systems and explainability techniques that faithfully explicit the behaviour of complex machine learning models tailoring their explanation in an understandable way for humans.

2.2. An Overview of the State of the Art

Given the broadness of the current state-of-the-art in Explainability and Explainable AI and the target of this article, we provide an overview of explainability methods to outline the variety of approaches available in the literature. For a complete and detailed summary of the state-of-the-art explainability, we advise the reader to refer to [2,9,10,17,18].

One of the most interesting intuitions conceived in this research field is that inherently explainable models can be employed to approximate the behaviour of black-box models. Such approximations can explain the original model as they are promptly understandable. One of the most well-known methods in this category is Local Interpretable Model-agnostic Explanations' (LIME) [19]. This post-hoc, model-agnostic, local explainability approach faithfully explains the predictions of any classifier or regressor by approximating it locally with an interpretable representation model. It was also extended to address the so-called "trusting the model" problem by developing Submodular Pick-LIME (SP-LIME) to explain multiple non-redundant instances of a model prediction. This process aims to increase users' trust in the whole model since providing and end-user with a single understandable outcome is not enough to achieve such an objective. Lundberg et al. [20] presented a unified framework for interpreting predictions named SHapley Additive exPlanations (SHAP). SHAP unifies six different local methods—including LIME [19] and DeepLIFT [21]—by defining the so-called class of additive feature attribution methods, described using the novel perspective that any explanation of a model's prediction is a model itself. The authors also present SHAP values as a unified measure of feature importance, propose a new estimation method and demonstrate that the computed values are better aligned with human intuition and discriminate better among model output classes.

An intuitive and straightforward way of explaining the local behaviour of a machine learning system is highlighting the different parts of the output considered by the model to make its prediction. This explanation format is generically referred to as highlight. It has been widely applied to explain the behaviour of models performing various tasks involving pictures (e.g., image classification, object detection, etc.) In this case, a highlight—usually a heatmap or saliency map overlapped on the considered picture—identifies the different pixels or groups of pixels the model considered to make its prediction. One of the best-known approaches employing such a format is Gradient-weighted Class Activation Mapping (Grad-CAM) [22]. It generates explanations of Convolutional Neural Network (CNN)-based models

by using the gradients of a concept of interest at the final convolutional layer to produce a localisation map highlighting the significant regions in the image for predicting the concept. The authors also described an extension named Guided Grad-CAM to create high-resolution class-discriminative visualisations with the ability to show fine-grained importance about the entity identified by combining Guided Backpropagation and Grad-CAM visualisations via pointwise multiplication. Despite outperforming state-of-the-art methods on both interpretability and faithfulness, Grad-CAM had some limitations, namely, a performance decrease when localizing multiple instances of the same class and the lack of completeness in identifying entities in single object images. Seeking to overcome them, Chattopadhyay et al. [23] proposed Grad-CAM++, enhancing Grad-CAM by improving object localisation and explaining multiple object instances in a single picture. Moreover, Grad-CAM++ was combined with SmootGrad [24] to strengthen its capabilities. Smooth Grad-CAM++ [25] improves object localisation even further by applying a smoothening [24] technique when computing the gradients involved in Grad-Cam++. It also provides visualisation capabilities to generate explanations for any layer, subset of feature maps or subset of neurons within a feature map at each instance at the inference level.

Highlights are also employed to shape the explanations of models performing Natural Language Processing (NLP) tasks (e.g., question answering, sentiment analysis, etc.) In this context, a highlight—usually represented as a saliency map between couples of words or saliency highlights, i.e., coloured boxes with varying colour intensities depending on the word relevance, overlapped to the input text—specifies the terms or the piece of text that the model employed to define the outcome of the task. Ghaeini et al. [26] utilised saliency visualisations to explain a neural model performing Natural Language Inference (NLI). Such a task requires the model to define the logical relationship between a premise and a hypothesis choosing between entailment, neutral or contradiction. The authors proposed and demonstrated the effectiveness of saliency maps in describing model behaviour on different inputs and between different models, revealing interesting insights and identifying the critical information contributing to the model decisions. Dunn et al. [27] combined dependency parsing, BERT [28], and the leave-n-out technique to develop a context-aware visualisation method leveraging existing NLP tools to find the groups of words that have the most significant effect on the output. It employs dependency parsing tools combined with a model-agnostic Leave-N-Out pre-processing to identify contextual groups of tokens that have the largest perceived effect on the model's classification output. Such a methodology produces saliency highlights with more relevant information about the classification and more accurate highlights.

Other techniques provide humans with examples, i.e., representative data samples to explain the model's behaviour. Kim et al. [29] introduced Concept Activation Vectors (CAVs) to interpret the internal state of a Neural Network (NN) in terms of human-friendly concepts. They employed the Testing with CAV (TCAV) technique to quantify how important a user-defined concept is to an image classification result. In particular, this methodology orders the set of pictures associated with any user-defined concept received in input based on the computed values. Jeyakumar et al. [15] described ExMatchina. This open-source explanation-by-example implementation identifies and provides the nearest matching data samples from the training dataset as representative examples applying cosine similarity. They also proved that users prefer this type of explanation for most tasks while still acknowledging that the main limitation of their method is the quality of the training data.

In conclusion, the literature in Explainable AI presents a wide variety of methods, principles and structures useful to collect insights about the behaviour of AI and ML systems. Such approaches are organized depending on their applicability, their characteristics and the explainability-related aspect they address. We presented the definitions we argue to be the most relevant ones and surveyed the literature to provide an overview of the variety of the available methods.

3. Research Methodology

Given the broadness of the literature on Explainability and Explainable AI and the impact of such a research field over the last years, we focus on articles and papers published over the last five years, from 2017 to 2022. We collected articles from bibliographic databases, combining input from both open-access (i.e., Google Scholar) and subscribers-only (i.e., Scopus) sources in the field of computer science. We implemented a strategy that is aligned with the PRISMA methodology [30] for literature reviews. We defined a search strategy to collect articles that include any pair of concepts created by combining the keywords listed in Table 1. In particular, all the possible pairs of keywords have been generated, by concatenating one explainability keyword (left column in the table) with one knowledge-related keyword (right column in the table). The reader can refer to Appendix A for the detailed structure of the queries performed.

Table 1. The list of keywords used to generate the couples used to search for papers.

Explainability-Related Keywords	Knowledge-Related Keywords
Interpretable Machine Learning	Knowledge Extraction
Explainable Machine Learning	Knowledge Elicitation
Explainable Artificial Intelligence	Crowdsourcing
Explainable AI	Human-in-the-Loop
Explainability	Human-centred Computing
Interpretability	Human-centred Computing
	Human Computation
	Concept Extraction

When querying Google Scholar, we excluded all the articles whose title contained the words "survey" and "review" while considering only the first 100 articles ranked by relevance for each query. We restricted our research to the top 100 results as we noticed a drop in pertinence to our topic of interest after the 80th position in the ranking. Notice that we do not have full control on the implementation of the search strategies run by the bibliographic databases: for instance, while Google runs its matching over the full text of the article, others may only search the metadata of the articles (title, abstract, keywords, categories, etc.)

In total, we examined: *(i)* 3718 non-unique articles from Google Scholar, extracted by querying the bibliographic database using 48 combinations of keywords performed through the tool *Publish or Perish 8.2.3944.8118* (Harzing, A.W. (2007) Publish or Perish, available from https://harzing.com/resources/publish-or-perish, accessed on 28 April 2022), finally resulting in 2056 unique papers; and *(ii)* and 327 non-unique articles from Scopus, queried using the Scopus web interface and following the same query criteria used for Google Scholar, resulting in 216 unique articles. Indeed, most of the queries performed on the Scopus bibliographic database returned very few results, as the scope of each of them was quite narrow.

By combining the two sources, we finally obtained an integrated set of 2197 unique articles to analyse. The authors manually inspected the collected articles, considering only the ones attaining to the scope of this review. In particular, we considered all and only the articles in which humans and human knowledge played a fundamental role in with respect to the explainability of the system. With the aim of collecting a broad selection of documents, no further exclusion condition was applied. Finally, the bibliographic references cited in the collected documents have also been inspected, consequently extending the considered literature.

4. Human Knowledge and Explainability

Bridging the understandability gap between humans and black-box models requires the development of techniques able to answer the many-faceted problem of explainability, addressing the faithfulness and completeness of the explanations representing the model's

behaviour, while also accounting for the capability of the human interpreter to understand it. In the field of machine learning, humans are commonly employed to collect or label data, debugging and evaluate the outcomes of machine learning models, and many more [31]. Due to the recent enthusiasm in XAI, researchers' data interests shifted towards collecting human knowledge in the form of human rationale [32], i.e., the reasoning applied by humans to perform a ML task. Such valuable [33] information is at the centre of many explainability-related tasks and can be employed in a wide variety of ways (summarised in Figure 1). In a broader sense, human knowledge is also (indirectly) applied in most human-in-the-loop approaches in which explanations are used as a means to explore [34], evaluate, or improve the explainability and (sometimes) the performance of models. Furthermore, humans are directly involved in the creation [35], assessment, or improvement [36] of such explanations or the model itself [37]. Given the critical role of human knowledge in such processes, keeping the human-in-the-loop is essential to achieve interpretable and explainable AI [38,39]. In the following sections, we report on a wide variety of approaches using humans and their knowledge to achieve such objectives and discuss their findings.

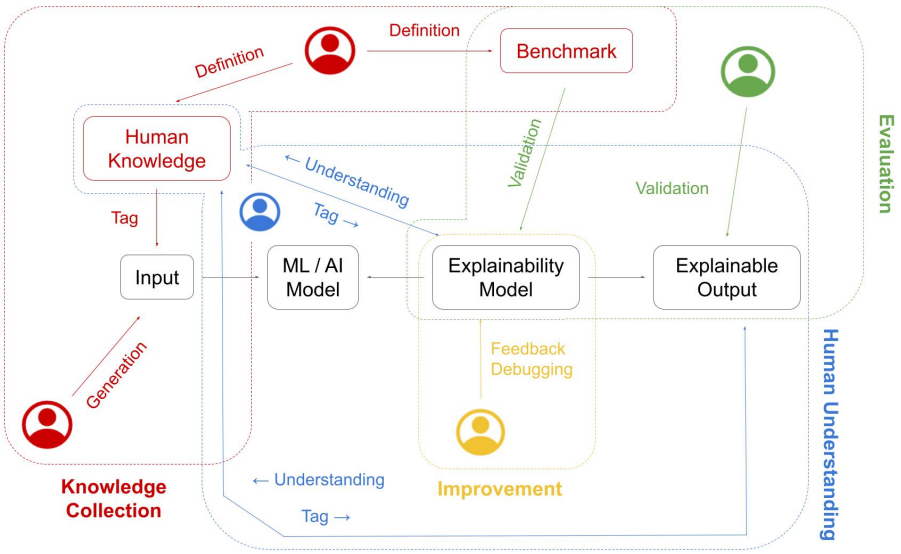

Figure 1. In the figure, the four main ways to use of human knowledge in explainability are represented, namely, knowledge collection for explainability (red), explainability evaluation (green), understanding human's perspective in explainability (blue), and improving model explainability (yellow). In the schema, the icons represent human actors.

4.1. Explainability and Human Knowledge Collection

In computer science, crowdsourcing is a well-known practice widely employed to collect a large amount of human-generated data by engaging heterogeneous groups of people with varying features and knowledge in undertaking a task [40]. Given the fundamental role of humans in XAI, crowd knowledge collection is fundamental to leverage human intelligence at scale to achieve robust, interpretable, and hence trustworthy AI systems [41]. When addressing the explainability of black-box models, many different factors influence such an approach. In particular, depending on the complexity of the system [42], the model's purpose, the complexity of the task, and its goal, it might be necessary to involve individuals with specific knowledge or features [43]. Indeed, complex explainability-related tasks may require preliminary expertise which translates into the involvement of expert users [44]. For example, collecting and employing human knowledge to label [45,46] and evaluate visual explanations (e.g., heatmaps) extracted from an image classification model could be trivial as users may be asked to just highlight the various parts of the picture they deem to be important [47]. On the other hand, editing attention maps to successfully improve the explainability of a system [48] or providing domain-specific

knowledge [32] are not tasks that can be easily accomplished by non-expert users. As a consequence, interactive approaches have been developed to employ human knowledge at its best while accounting for such complexities.

In the context of Question Answering systems, Li et al. [49] collected a dataset by engaging crowdworkers in interacting with the system and providing feedback on the quality of the answers both in a structured and unstructured way. The collected data were then employed to train a new model extending the original one with re-scoring and explanation capabilities. In the context of image classification tasks, Mishra et al. [47] designed a concept elicitation pipeline to gather high-level concepts to build explanations for image classification datasets. The data were collected as mask–label pairs by showing the picture's true label to users and having them outline both the entity and some of the features they used to identify it. Per-image and per-class aggregations were employed to build a variety of concept-driven explanations. Similarly, Uchida et al. [50] proposed a human-in-the-loop approach collecting human knowledge to generate logical decision rules to explain the output of classification models. They explained the outcome of the original model by collecting human-interpretable features of pictures as text to generate rule tables associating the classes and the collected features. Balayn et al. [51,52] proposed a Game With a Purpose (GWAP) to collect high-quality discriminative and negative knowledge. Inspired by the popular game *GuessWho?*, users are engaged in a competitive, two-player game in which each user should guess the card chosen by the challenger by asking questions about the entity represented on the card. The answers to such questions represent the (structured) knowledge collected about the entity of choice. Such a particular kind of knowledge can be useful to improve the trustworthiness and robustness of AI systems. While [47,50] directly involved users in the description of a series of pictures, Tocchetti et al. [53] proposed a two-player gamified activity to collect human knowledge describing different features of real-world entities while unbinding pictures from the feature description process. Indeed, one of the players is asked to predict the entity in the picture by guessing its features through closed questions while the other player provides the answers, classifies and outlines the guessed features on the image. The described methods generate explanations and/or collect features while unbinding the model itself from the data collection process, using only its input. While such an approach eases the data collection process, employing the explanations of a model enhances and contextualises the collected content. Zhao et al. [54] designed ConceptExtract, a system implementing a human-in-the-loop approach to generate user-defined concepts for Deep Neural Network (DNN) interpretation. Users can overview and filter image patches extracted from the input pictures, provide new visual concepts, and overview the performance and the interpretation of the target model. Attempting to achieve a similar objective, Lage et al. [55] proposed a human-in-the-loop approach to learn a set of transparent concept definitions relying on the labelling of concept features. Users were engaged to provide their understanding of the domain of interest, consequently making the collected concepts intuitive and interpteable. In particular, they were asked to define the associations between a series of features and concepts, and provide feedback on whether the function learned by the model satisfies the aforementioned conditions. A process similar to [54] was employed in the development of FaxPlainAC [56], a tool to collect user feedback on the outcome of explainable fact-checking models. When a query is received by the system, its decision, i.e., the truthfulness of the input fact, and the considered evidence are displayed. Users are asked whether the documents employed by the system to generate such content are supporting or refuting the input by highlighting the most relevant parts of the text, or whether they are misleading or irrelevant. Sevastjanova et al. [57] extended the usage of explainability to support interactive data labelling of complex classification tasks by applying visual-interactive labelling and gamification. Such an approach is implemented in QuestionComb, a rule-based learning model that presents explanations as rules, supporting iterative and interactive optimisation of the data. These methods demonstrate that data collection processes may employ explanations and model details to improve the level of

detail and accuracy of the collected knowledge. Furthermore, the strategy to apply is also influenced by the kind of data desired, i.e., task-specific or generic, resulting in the design of a variety of data collection techniques.

4.2. Evaluation of Explainability Methods by Means of Human Knowledge

The design and implementation of approaches to choose the best explainability method or explainable model has been at the centre of discussion of the research community for years. Consequently, recent research efforts have focused on the collection and development of benchmarks due to their capability to enable, organise and standardise the evaluation and comparison of multiple models by means of explainability-related measures. Mohseni et al. [58] developed a benchmark for quantitative evaluation of saliency map explanations of images and text tasks through multilayer, aggregated human attention masks. They collected human annotations of salient features by asking users to highlight the most representative parts of documents or images. The efficacy of their approach was validated through a series of experiments, demonstrating its capabilities to evaluate the completeness and correctness of model saliency maps. De Young et al. [59] proposed the Evaluating Rationales And Simple English Reasoning (ERASER) benchmark, comprising various datasets and tasks extended with human annotations of rationale. Such datasets cover various NLP tasks, such as question answering, sentiment analysis, etc. They evaluated their benchmark on a set of baseline models with respect to a set of proposed metrics designed to measure faithfulness and the agreement between human annotations and model's extracted rationales. While benchmarks provide fixed datasets to evaluate model explainability, Schuessler et al. [60] developed a library that allows researchers to create customized datasets for human-subject and algorithmic evaluations of explanation techniques for image classification.

The employment of automatic metrics to evaluate and compare model explainability is still an interesting topic of debate and interest within the XAI literature. In particular, it is argued that the metrics used to evaluate explainability methods must be chosen carefully while there is significant room for improvement for such assessment approaches [61]. Moreover, exploring the relation between human-based and automatic evaluations is another aspect researched in the XAI community [62]. On such a topic, while a variety of evaluation methods and approaches have been proposed [63], it is still argued that the best way to assess the interpretability of black-box models is through user experiments and user-centred evaluations as there is no guarantee for the correctness of automated metrics in evaluating explainability [64] and high explainability metric scores do not necessarily reflect high human interpretability in real-world scenarios [64,65]. The same is true for well-known metrics (e.g., F1-score) [66]. Supporting such claims, Fel et al. [65] conducted experiments to evaluate the capability of human participants to leverage representative attribution methods to learn to predict the decision of various image classifiers. Such a process was aimed at assessing the usefulness of explainability methods and the capability of existing theoretical measures in predicting their usefulness in practice. The framework they designed can be employed to perform such an evaluation given a black-box model, an explanation method and a human subject to predict the predictor (i.e., the so-called meta-predictor). A two phase procedure is applied. In the learning phase, the human meta-predictor is trained using triples made of an input sample, the model's prediction and its explanation to uncover rules describing the functioning of the model. In the evaluation phase, the accuracy of the meta-predictor—and consequently, the relevance of the rules they learned—is tested on new samples by comparing their predictions with the ones provided by the model. In their conclusions, the authors argue that faithfulness evaluations are poor substitutes for utility and it is necessary to put the human in the loop. Moreover, they discuss that such metrics do not account for the usefulness of the explanation to humans as in some cases they can either be not useful or generate ambiguity. We argue that the main problem is not related to the application of automatic evaluations and metrics, but on the interpretation of the computed (faithfulness) scores. Faithfulness is just one side of

the coin, i.e., the model's side, as it measures how close the derived explanation is with respect to the true reasoning process of the model. The other side of the coin is represented by interpretability, i.e., a human interpreter should be able to properly understand the explanation. The misunderstanding occurs when there is confusion between these two aspects. Indeed, model faithfulness and interpretability are not to be considered equivalent when it comes to the evaluation of the explainability of models.

The evaluation of the interpretability of the explanations of a black-box model is usually performed by involving users in manually interpreting the explanations generated by the system or derived through explainability methods. The same approach is applicable to the evaluation of the interpretability of black-box models, i.e., directly understanding the intrinsic explainability of a model [67]. Such evaluations are usually achieved through user questionnaires [66,68–70] whose questions vary depending on the nature of the experiment, model, etc. On the other hand, comparing the interpretability of different explainability methods to choose the best suited one requires the design and implementation of ad hoc human-in-the-loop approaches. Soltani et al. [71] improved existing XAI algorithm by employing cognitive theory principles with the final aim of providing explanations similar to domain experts. Humans were involved in a series of experiments aimed at evaluating both the novel approach and the basic one to understand which one led to the best explanations. In their work, Lu et al. [64] designed a novel human-based evaluation approach using crowdsourcing to evaluate saliency-based XAI methods—mainly focusing on methods that explain the prediction of picture-based models, e.g., Grad-CAM [22], SmoothGrad [24], etc.—through a human computation game named "Peek-a-boom". Their human-centred approach compares different Explainable AI methods to identify the one yielding to the best interpretations. In the proposed Game With a Purpose (GWAP), the XAI method plays the role of Boom, revealing parts of an image as the game progresses, and the player plays the role of Peek, guessing the entity in the picture from the parts displayed. In summary, evaluating the explainability of black-box models requires assessing both human interpretability and faithfulness, while not misunderstanding these two concepts and consequently generating unmotivated trust.

More commonly, humans are engaged to evaluate the effectiveness of methods in generating explanations and their usefulness in real scenarios [72–75]. Zhao et al. [73] employed Generative Adversarial Networks (GANs) to generate counterfactual visual explanations. Crowd workers were recruited to evaluate their effectiveness for classification. In the context of Visual Question Answering, Arijit et al. [74] involved users in a collaborative image retrieval game, named Explanation-assisted Guess Which (ExAG), to evaluate the efficacy of explanations, finally demonstrating the usefulness of explanations in their setting. Alvarez-Melis et al. [75] implemented a method to generate explanations based on the concept of weight of evidence from information theory. User experiments demonstrated the effectiveness of the methodology in generating accurate and robust explanations, even in high-dimensional, multi-class settings. Zeng et al. [76] present a human-in-the-loop approach to explain ML models using verbatim neighbourhood manifestation. A three-stage process is employed to (i) generate instances based on the chosen sample, (ii) classify the generated instances to define the local decision boundary and delineate the model behaviour, and (iii) involve users in refining and explore the neighbourhood of interest. A series of experiments revealed the effectiveness of the implemented tool in improving human understanding of model behaviour. Baur et al. [77] presented NOVA, a human-in-the-loop annotation tool to interactively train classification models from annotated data. The tool allows the employment of semi-supervised active learning to pre-label data automatically. Moreover, it implements recent XAI techniques to provide users with a confidence value of the predicted annotations and visual explanations. Heimerl et al. [69] employed NOVA in emotional behaviour analysis. They engaged non-expert users and evaluated the impact and the quality of the explanations extracted, revealing their effectiveness in the presented use-case while getting useful insights on the employment of visual explanations. Steging et al. [32] proposed a knowledge-driven method for model-agnostic

rationale evaluation employing human-in-the-loop to collect dedicated test sets to evaluate targeted rationale elements based on expert knowledge of the domain.

Finally, while part of the XAI research community focused on designing and implementing methods to generate explanations, the development of techniques aimed at generating trust in models is another fundamental aspect of interest. Zöller et al. [78] implemented XAutoML, an interactive visual analytic tool aimed at estabilishing trust in AutoML-generated models. The user-centered experiments reveled the effectiveness of the tool in generating trust while addressing the explainability needs of various user groups (i.e., domain experts, data scientists, and AutoML researchers). De Bie et al. [79] proposed and evaluated RETRO-VIZ, a method to estimate and evaluate trustworthiness of regression prediction. The system comprises RETRO, a method to quantitatively estimate the trustworthiness of the prediction, and VIZ, a visualisation provided to users to identify the reasons for the estimated trustworthiness. Although they demonstrated the effectiveness of their methodology, the authors remark it must be used with caution as to not generate unguided trust.

4.3. Understanding the Human's Perspective in Explainable AI

An explanation that cannot be properly understood by a human has no value and may potentially mislead the user. Indeed, it is essential to provide accurate and understandable explanations as poor explanations can sometimes be even worse than no explanation at all [80] and may also generate undesired bias in the users [81,82]. As a consequence, properly structuring [83] and evaluating the interpretability and effectiveness of explanations requires a deep understanding of the ways in which humans interpret and understand them, while also accounting for the relationship between human understanding and model explanations [84,85]. For such reasons, the explainable AI research field spreads from IT-related fields, such as computer science and machine learning, to a variety of human-centred disciplines, such as psychology, philosophy, and decision making [86]. Therefore, recent studies aimed at evaluating human behaviours when exploring, interpreting and using explanations have been conducted [12,13,87]. Moreover, Gamification and Games With a Purpose have been proven to be quite effective in assessing how humans interpret XAI explanations [88]. Feng et al. [6] evaluated how humans employ model interpretations and their effectiveness, measured in terms of improvement in human performance. They designed Quizbowl, a human–computer cooperative setting for question answering, supporting various forms of interpretations whose objective is to guide users to decide whether to trust the model's prediction or not. The question to answer is displayed word-by-word and players are asked to stop the display as soon as the model's interpretations are enough to answer the question correctly, but before it is completely revealed. They discovered that interpretations help both non-expert users and experts in different ways. Additionally, while expert users were able to mentally tune out bad suggestions, novice users trusted the model too much, consequently choosing an incorrect answer. Such a result demonstrates that even though one of the objectives of explainability is to improve the user's trust in the model, it is necessary to organise the content provided as to avoid generating a sense of overconfidence in the system. A similar result was achieved by Ghai et al. [89], who combined XAI techniques in the context of Active Learning. They analysed the impact of the proposed approach, while also researching on human-related aspects. Their findings revealed that explanations successfully supported users with high task knowledge, while impairing those with low task knowledge. Indeed, users with low knowledge were more prone to agreeing with the model, even when it misbehaved. On the other hand, they were able to demonstrate the effectiveness of explanations in calibrating user trust and evaluating the maturity of the model. In conclusion, achieving a high level of transparency is not always beneficial to improving the user's understanding [5,81]. Indeed, providing complex or a large number of explanations would generate a trade-off between their understandability and the time required by human interpreters to interpret them [42,90]. Consequently, it is necessary to comprehend the proper level of transparency, explanation complexity and

quantity, even in simple cases [91]. Regarding such an aspect, Mishra et al. [47] performed user studies to understand the proper level of conceptual mapping by means of granularity and context of the data to generate explanations. The authors discovered that a balance between coarse and fine-grained explanations help users understand and predict the model's behaviour. On the contrary, the usage of structured coarse-grained explanations negatively impacted user's trust and performance. While Mishra et al. [47] focused on understanding the granularity of the explanations, Kumar et al. [92] compared the visual explanations provided by the proposed visualisation framework with respect to two text-based baselines, revealing the effectiveness of their approach in the context of interest through user experiments. In conclusion, engaging humans in XAI is fundamental, as they are the target of the explanations and improving our understanding of their behaviour when interacting with explanations and models is beneficial to improving the design and development of explanations. Furthermore, it is desirable to design flexible explanation approaches and explainability methods able to properly convey model behaviour depending on "who" the human is [91,93,94]. A categorisation of the main user groups is provided by Turró [93]. Depending on their *goals, background and relationship with the product*, users are grouped in three categories: developers and AI researchers, domain experts and lay users. The author discusses the importance of approaching explainable AI in a user-centered manner, providing tailored explanations based on the needs and characteristics of the targeted group of users, finally improving affordability and user satisfaction, and easing the explanation evaluation process. Striving to understand how and why such groups employ explanations and behave, several researchers have carried out experiments by engaging specific user groups. Hohman et al. [95] involved professional data scientists to explore how and why they interpret ML models and how explanations can support answering interpretability-related questions. More generally, users can be classified as domain or expert users and non-expert users. Nourani et al. [96] inspected the behaviour of such user groups on their first impression of an image classification model based on the correctness of its predictions. They discovered that providing early errors to domain experts decreases their trust, while early correct predictions help them in adjusting their trust based on their observations of the system performance. On the other hand, non-expert users relied too much on the predictions made by the model due to their lack of knowledge. Such over-reliance on the ML system [6,89,96] highlights how it is always necessary to account for the users engaged in the system. Moreover, while it is necessary to engage non-expert and end users in the evaluation of such system, it is also recommended to consider their features, preliminary knowledge and understanding of the system of interest.

Finally, while explanations were proven to be effective in leading the user in achieving a task and improving their trust and understanding of the model, it has also been demonstrated that sometimes they are either not able to improve [97,98] or, worse, they reduce human accuracy and trust [99]. A similar result in a different context was found by Dinu et al. [100]. They focused on post hoc feature attribution explanations and discovered that such explanations *provide marginal utility in our task for a human decision maker and, in certain cases, result in worse decisions due to cognitive and contextual confounders*. Such findings bring forth a fundamental conclusion. Even though explanations and explainability methods may improve users' understanding, accuracy and trust [101], it is still necessary to investigate the way humans perceive such content with respect to the context, the model and the task it performs.

4.4. Human Knowledge as a Mean to Improve Explanations

As faithful explanations provide meaningful insights into the behaviour of models, researchers have designed novel and effective methods to employ such content to improve the explainability and performance of models. Such human-in-the-loop approaches mainly display the explanations and the outcomes of a model to humans who are then asked to discover undesired behaviours (i.e., debugging the model) and to provide possible corrections. The effectiveness of such explainability-focused approaches is discussed by Ferguson et al. [102] They

report on the usefulness of explanations for human–machine interaction, while stating that augmenting explanations to support human interaction enhances their utility, creating a common ground for meaningful human–machine collaboration. They experienced the effectiveness of editable explanations, consequently modifying the machine learning system to adapt its behaviour to produce interpretable interfaces. Many examples of approaches that make use of such a strategy can be found in the literature. Mitsuhara et al. [48] proposes a novel framework to optimise and improve the explainability of models by using a fine-tuning method to embed human knowledge—collected as single-channel attention maps manually edited by human experts—in the system. They reveal that improving the model's explainability also contributes to a performance improvement. Coma et al. [103] designed an iterative, human-in-the-loop approach aimed at improving both the performance and the explainability of a supervised model detecting non-technical losses. In particular, each iteration improves (or at least does not deteriorate) the performance and reduces the complexity of the model to improve its interpretability. Kouvela et al. [104] implemented Bot-Detective, a novel explainable bot-detection service offering interpretable, responsible AI-driven bot identification, focused on efficient detection and interpretability of the results. Users can provide feedback on the estimated score and the quality of the results' interpretation, while specifying their agreement and describing eventual improvements of the explanations provided through LIME [19]. Such an approach not only improves the explainability of the model, but also contributes to the performance of the model itself. Collaris et al. [105] introduced an interactive explanation system to explore, tune and improve model explanations. The tool allows stakeholders to tune explanation-related parameters to meet their preferences while they employ such evidence to diagnose the model and discover eventual model or explanation improvements. Yang et al. [106] addresses the problem of generalisability by allowing users to co-create and interact with the model. The authors introduced RulesLearner, a tool able to express ML models as rules, while allowing users to interact with and update the patterns learned. Their studies demonstrated the effectiveness of the proposed approach in improving the generalisability of the analysed system and the quality of the explanations employed in the process. In the presented systems, users directly interact with the explanations of the model to improve their explainability. Other studies collect and employ human rationales [107] or domain knowledge [108,109] to achieve the same goal. Arous et al. [107] introduced MARTA, a Bayesian framework for explainable text classification. Such a system integrates human rationales into attention-based models to improve their explainability. Confalonieri et al. [108] evaluated how ontologies can be used to improve the human understandability of global post hoc explanations, presented as decision trees. The proposed algorithm enhances the explanation extracted using domain knowledge modelled as ontologies. While sometimes increasing the performance of the model is a side effect of improving its explainability [48,103–105], a few researchers employed explanations as a means to directly improve model performance [49,110,111]. Li et al. [49] collected human feedback, made of a rating label and a textual explanation describing the quality of the answer, to improve the performance and the capability of explaining the correctness of the outcome of a BERT-based Question Answering model. While [49] employed human feedback, Spinner et al. [111] engaged humans in a conceptual framework focused on practicability, completeness and full coverage to operationalise interactive and explainable machine learning. The most relevant element of the system is the Explainable AI pipeline which maps the explainability process to an iterative workflow that allows users to understand and diagnose the system to refine and optimise the model. Differently from the methods presented, Hadash et al. [112] did not design a human-in-the-loop approach. Instead, they applied "positive framing" and improved "semantic labelling" to explanations—extracted through SHAP [20] or LIME [19]—to enhance model-agnostic explainable AI tools.

Another process benefiting from faithful explanations is model debugging. Such an activity employs human knowledge and expertise to identify errors, bias and improper behaviours in models with the final objective of correcting them, consequently improving the model [113] and/or its explanations. The main concept on which model debugging is based is the interactive exploration of models [80,114,115] by means of an interface able to

summarise its behaviour. Moreover, allowing users to interact with explanations produce an even deeper understanding of the model behaviour, consequently improving their capability to identify potential bugs. In this specific scenario, providing faithful, complete and understandable explanations is extremely important as they influence the capability of users of identifying such errors and the soundness of the results. With the final aim of understanding the model's failures, Nushi et al. [116] implemented Pandora, a system leveraging human and system-generated observations to describe and explain when and how ML systems fail. The tool employs content-based views (i.e., views creating a mapping between input and the overall system's failure) to explain when the system fails while component-based views (i.e., views modelling how internal model dynamics lead to errors) explain how the system fails. Crowdsourced human knowledge is employed for a variety of purposes, such as system evaluation, content data collection and component quality features data collection. Liu et al. [117] describe an error detection framework for sentiment analysis models based on explainable features employing a variety of explanations. Their approach is organised in four different units, namely, a "local-level feature contributions" module extracting unigram features through LIME [19], a "global-level feature contributions" module performing perturbation-based analyses by masking individual features of the training samples, a "human assessment" module asking humans to assess the most relevant globally contributing features learned from the previous step, and a "global-local integration" module that quantifies the erroneous probabilities of instance-level predictions made by the model. Even though providing a wide variety of interactive explanations may contribute to improving the debugging of ML systems, it is still unclear which ones are the most useful. Seeking to answer such a question, Balayn et al. [118] developed an interactive design probe that provides various explainability functionalities in the context of image classification models. They discovered that common explanations are primarily used due to their simplicity and familiarity while other types of explanation, e.g., domain knowledge, global, textual, active, interactive, and binary explanations are still useful to achieve a variety of objectives. Such conclusions support and highlight the importance of presenting diverse explanations. Using explanations as a means to debug models could also benefit the explanations themselves. For example, Afzal et al. [119] described a human-in-the-loop explainability framework to debug data issues to enhance interpretability and facilitate informed remediation actions. In conclusion, the variety of human-in-the-loop approaches presented demonstrates that human knowledge can be a valuable asset even for tasks that do not employ it as structured data and directly engage humans in the process of understanding, fixing and optimizing ML models.

5. Conclusions

In this article, we presented an overview of the last five years of literature about explainability and Explainable AI, framed from the human perspective and focused on human-in-the-loop approaches and techniques employing human knowledge to achieve their goals. We argue that human knowledge is not necessarily associated with the notion of data, but also with the capability of humans to accomplish tasks and the reasoning they apply i.e., human rationale. We cover explainability-related topics employing such knowledge in a wide variety of ways, e.g., training data, explainability evaluation, model and explainability improvement, etc. We argue that humans and their knowledge play a fundamental role in the field of Explainable AI. In particular, improving and assessing the interpretability of models is a task requiring active human involvement. The same is true for model debugging. Recent studies have focused on the human's side of explainability, focusing on comprehending how to shape explanations to make them more interpretable and how humans employ and understand them. Such studies are of fundamental importance in this research field, as humans are not (only) "data sources" for our models, but also the targets of the explanations and the models we strive to improve and refine.

Many questions are yet to be answered in this research field. We argue that one of the most fundamental and complex ones is the proper way of structuring explanations with

respect to the users and the context. Moreover, while a variety of explainability methods and models are available in the literature, the choice of which one to employ is still at the centre of discussion. Answering such questions requires accounting for the intrinsic complexity of humans and the context in which they are put. In conclusion, we argue that humans and their knowledge are both the reason for the existence of this research field and the solution to many of the complex questions under active research. Future research in the field of Explainable AI and Explainability should focus their efforts on developing heuristics and methods to (1) properly evaluate and compare model explainability, i.e., able to consider a variety of aspects both related with models and humans (e.g., faithfulness and interpretability), (2) design generalisable methods able to deal with a wide variety of contexts and models, and (3) explore the intrinsic complexity associated with humans' and models' contexts.

Funding: This research is partially supported by the European Commission under the H2020 framework, within project 822735 TRIGGER (TRends in Global Governance and Europe's Role) and by the contribution of the Ph.D. Scholarship on Explainable AI funded by Cefriel.

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.

Appendix A

In this part of the appendix, we describe the queries performed on the different bibliographic databases analysed.

The queries performed on Scopus abide by the following structure:
TITLE-ABS-KEY ("Explainability-related Keyword" AND "Knowledge-related Keyword") AND LIMIT-TO(SUBJAREA , "COMP") AND PUBYEAR > 2016.

The queries performed on Google Scholar through *Publish or Perish 8.2.3944.8118* were performed by properly compiling the dedicated fields in the application, namely

- **Years**: *2017–2022*;
- **Title words**: *NOT Survey NOT Review NOT Systematic*;
- **Keywords**: *"Explainability-related Keyword", "Knowledge-related Keyword"*.

All the pairs containing an *Explainability-related Keyword* and *Knowledge-related Keyword* were queried.

References

1. Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P.; Cao, X.; Khosravi, A.; Acharya, U.R.; et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf. Fusion* **2021**, *76*, 243–297. [CrossRef]
2. Vilone, G.; Longo, L. Explainable Artificial Intelligence: A Systematic Review. *arXiv* **2020**, arXiv:2006.00093.
3. Chou, Y.; Moreira, C.; Bruza, P.; Ouyang, C.; Jorge, J.A. Counterfactuals and Causability in Explainable Artificial Intelligence: Theory, Algorithms, and Applications. *Inf. Fusion* **2022**, *81*, 59–83. [CrossRef]
4. Holm, E.A. In defense of the black box. *Science* **2019**, *364*, 26–27. [CrossRef]
5. Poursabzi-Sangdeh, F.; Goldstein, D.G.; Hofman, J.M.; Vaughan, J.W.; Wallach, H.M. Manipulating and Measuring Model Interpretability. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021.
6. Feng, S.; Boyd-Graber, J.L. What can AI do for me: Evaluating Machine Learning Interpretations in Cooperative Play. In Proceedings of the 24th International Conference on Intelligent User Interfaces, Marina del Ray, CA, USA, 17–20 March 2019.
7. Hahn, T.; Ebner-Priemer, U.; Meyer-Lindenberg, A. Transparent Artificial Intelligence—A Conceptual Framework for Evaluating AI-based Clinical Decision Support Systems. *OSF Preprints* **2019**. [CrossRef]
8. O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*; Crown Publishing Group: New York, NY, USA, 2016.
9. Guidotti, R.; Monreale, A.; Turini, F.; Pedreschi, D.; Giannotti, F. A Survey Of Methods For Explaining Black Box Models. *ACM Comput. Surv.* **2018**, *51*, 1–42. [CrossRef]

10. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
11. Narang, S.; Raffel, C.; Lee, K.; Roberts, A.; Fiedel, N.; Malkan, K. WT5?! Training Text-to-Text Models to Explain their Predictions. *arXiv* **2020**, arXiv:2004.14546.
12. Narayanan, M.; Chen, E.; He, J.; Kim, B.; Gershman, S.; Doshi-Velez, F. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *arXiv* **2018**, arXiv:1802.00682.
13. Xu, F.; Li, J.J.; Choi, E. How Do We Answer Complex Questions: Discourse Structure of Long-form Answers. *arXiv* **2022**, arXiv:2203.11048. [CrossRef]
14. Schuff, H.; Yang, H.; Adel, H.; Vu, N.T. Does External Knowledge Help Explainable Natural Language Inference? Automatic Evaluation vs. Human Ratings. *arXiv* **2021**, arXiv:2109.07833.
15. Jeyakumar, J.V.; Noor, J.; Cheng, Y.H.; Garcia, L.; Srivastava, M. How Can I Explain This to You? An Empirical Study of Deep Neural Network Explanation Methods. In Proceedings of the Advances in Neural Information Processing Systems, Online, 6–12 December 2020; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 4211–4222.
16. Sokol, K.; Flach, P.A. Explainability Is in the Mind of the Beholder: Establishing the Foundations of Explainable Artificial Intelligence. *arXiv* **2021**, arXiv:2112.14466.
17. Danilevsky, M.; Qian, K.; Aharonov, R.; Katsis, Y.; Kawas, B.; Sen, P. A Survey of the State of Explainable AI for Natural Language Processing. *arXiv* **2020**, arXiv:2010.00711.
18. Carvalho, D.; Pereira, E.; Cardoso, J. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* **2019**, *8*, 832. [CrossRef]
19. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv* **2016**, arXiv:1602.04938.
20. Lundberg, S.M.; Lee, S. A unified approach to interpreting model predictions. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
21. Shrikumar, A.; Greenside, P.; Kundaje, A. Learning Important Features Through Propagating Activation Differences. In Proceedings of the 34th International Conference on Machine Learning, PMLR, Sydney, NSW, Australia, 6–11 August 2017.
22. Selvaraju, R.R.; Das, A.; Vedantam, R.; Cogswell, M.; Parikh, D.; Batra, D. Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization. *arXiv* **2016**, arXiv:1611.07450.
23. Chattopadhyay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018.
24. Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.B.; Wattenberg, M. SmoothGrad: Removing noise by adding noise. *arXiv* **2017**, arXiv:1706.03825.
25. Omeiza, D.; Speakman, S.; Cintas, C.; Weldemariam, K. Smooth Grad-CAM++: An Enhanced Inference Level Visualization Technique for Deep Convolutional Neural Network Models. *arXiv* **2019**, arXiv:1908.01224.
26. Ghaeini, R.; Fern, X.; Tadepalli, P. Interpreting Recurrent and Attention-Based Neural Models: A Case Study on Natural Language Inference. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 4952–4957. [CrossRef]
27. Dunn, A.; Inkpen, D.; Andonie, R. Context-Sensitive Visualization of Deep Learning Natural Language Processing Models. In Proceedings of the 2021 25th International Conference Information Visualisation (IV), Sydney, Australia, 5–9 July 2021.
28. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
29. 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). *arXiv* **2017**, arXiv:1711.11279. [CrossRef]
30. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Syst. Rev.* **2021**, *10*, 89. [CrossRef] [PubMed]
31. Abhigna, B.S.; Soni, N.; Dixit, S. Crowdsourcing—A Step Towards Advanced Machine Learning. *Procedia Comput. Sci.* **2018**, *132*, 632–642. [CrossRef]
32. Steging, C.; Renooij, S.; Verheij, B. Discovering the Rationale of Decisions: Experiments on Aligning Learning and Reasoning. *arXiv* **2021**, arXiv:2105.06758.
33. Strout, J.; Zhang, Y.; Mooney, R.J. Do Human Rationales Improve Machine Explanations? *arXiv* **2019**, arXiv:1905.13714.
34. Gomez, O.; Holter, S.; Yuan, J.; Bertini, E. ViCE: Visual Counterfactual Explanations for Machine Learning Models. In Proceedings of the 25th International Conference on Intelligent User Interfaces, Cagliari, Italy, 17–20 March 2020.
35. Magister, L.C.; Kazhdan, D.; Singh, V.; Liò, P. GCExplainer: Human-in-the-Loop Concept-based Explanations for Graph Neural Networks. *arXiv* **2021**, arXiv:2107.11889.
36. Wang, J.; Zhao, C.; Xiang, J.; Uchino, K. Interactive Topic Model with Enhanced Interpretability. In Proceedings of the IUI Workshops, Los Angeles, CA, USA, 20 March 2019.

37. Lage, I.; Ross, A.S.; Kim, B.; Gershman, S.J.; Doshi-Velez, F. Human-in-the-Loop Interpretability Prior. *arXiv* **2018**, arXiv:1805.11571. [CrossRef]
38. Celino, I. Who is this explanation for? Human intelligence and knowledge graphs for eXplainable AI. In *Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges*; IOS Press: Amsterdam, The Netherlands, 2020.
39. Estivill-Castro, V.; Gilmore, E.; Hexel, R. Human-In-The-Loop Construction of Decision Tree Classifiers with Parallel Coordinates. In Proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada, 11–14 October 2020; pp. 3852–3859. [CrossRef]
40. Estellés-Arolas, E.; de Guevara, F.G.L. Towards an integrated crowdsourcing definition. *J. Inf. Sci.* **2012**, *38*, 189–200. [CrossRef]
41. Gadiraju, U.; Yang, J. What Can Crowd Computing Do for the Next Generation of AI Systems? In Proceedings of the CSW@NeurIPS, Online, 11 December 2020.
42. Lage, I.; Chen, E.; He, J.; Narayanan, M.; Kim, B.; Gershman, S.J.; Doshi-Velez, F. Human Evaluation of Models Built for Interpretability. *Proc. AAAI Conf. Hum. Comput. Crowdsourcing* **2019**, *7*, 59–67.
43. Lampathaki, F.; Agostinho, C.; Glikman, Y.; Sesana, M. Moving from 'black box' to 'glass box' Artificial Intelligence in Manufacturing with XMANAI. In Proceedings of the 2021 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC), Cardiff, UK, 21–23 June 2021; pp. 1–6. [CrossRef]
44. Hudec, M.; Mináriková, E.; Mesiar, R.; Saranti, A.; Holzinger, A. Classification by ordinal sums of conjunctive and disjunctive functions for explainable AI and interpretable machine learning solutions. *Knowl.-Based Syst.* **2021**, *220*, 106916. [CrossRef]
45. Sharifi Noorian, S.; Qiu, S.; Gadiraju, U.; Yang, J.; Bozzon, A. What Should You Know? A Human-In-the-Loop Approach to Unknown Unknowns Characterization in Image Recognition. In Proceedings of the ACM Web Conference 2022, Virtual Event, Lyon France, 25–29 April 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 882–892. [CrossRef]
46. Balayn, A.; Soilis, P.; Lofi, C.; Yang, J.; Bozzon, A. What Do You Mean? Interpreting Image Classification with Crowdsourced Concept Extraction and Analysis. In Proceedings of the Web Conference 2021, Ljubljana, Slovenia, 19–23 April 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 1937–1948. [CrossRef]
47. Mishra, S.; Rzeszotarski, J.M. Crowdsourcing and Evaluating Concept-Driven Explanations of Machine Learning Models. *Proc. ACM Hum.-Comput. Interact.* **2021**, *5*, 1–26. [CrossRef]
48. Mitsuhara, M.; Fukui, H.; Sakashita, Y.; Ogata, T.; Hirakawa, T.; Yamashita, T.; Fujiyoshi, H. Embedding Human Knowledge in Deep Neural Network via Attention Map. *arXiv* **2019**, arXiv:1905.03540.
49. Li, Z.; Sharma, P.; Lu, X.H.; Cheung, J.C.K.; Reddy, S. Using Interactive Feedback to Improve the Accuracy and Explainability of Question Answering Systems Post-Deployment. *arXiv* **2022**, arXiv:2204.03025. [CrossRef]
50. Uchida, H.; Matsubara, M.; Wakabayashi, K.; Morishima, A. Human-in-the-loop Approach towards Dual Process AI Decisions. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; pp. 3096–3098. [CrossRef]
51. Balayn, A.; He, G.; Hu, A.; Yang, J.; Gadiraju, U. FindItOut: A Multiplayer GWAP for Collecting Plural Knowledge. In Proceedings of the Ninth AAAI Conference on Human Computation and Crowdsourcing, Online, 14–18 November 2021; p. 190.
52. Balayn, A.; He, G.; Hu, A.; Yang, J.; Gadiraju, U. Ready Player One! Eliciting Diverse Knowledge Using A Configurable Game. In Proceedings of the ACM Web Conference Virtual Event, Lyon France, 25–29 April 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 1709–1719. [CrossRef]
53. Tocchetti, A.; Corti, L.; Brambilla, M.; Celino, I. EXP-Crowd: A Gamified Crowdsourcing Framework for Explainability. *Front. Artif. Intell.* **2022**, *5*, 826499. [CrossRef] [PubMed]
54. Zhao, Z.; Xu, P.; Scheidegger, C.; Ren, L. Human-in-the-loop Extraction of Interpretable Concepts in Deep Learning Models. *IEEE Trans. Vis. Comput. Graph.* **2022**, 28, 780–790. [CrossRef]
55. Lage, I.; Doshi-Velez, F. Learning Interpretable Concept-Based Models with Human Feedback. *arXiv* **2020**, arXiv:2012.02898.
56. Zhang, Z.; Rudra, K.; Anand, A. FaxPlainAC: A Fact-Checking Tool Based on EXPLAINable Models with HumAn Correction in the Loop. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Virtual Event, Queensland, Australia, 1–5 November 2021.
57. Sevastjanova, R.; Jentner, W.; Sperrle, F.; Kehlbeck, R.; Bernard, J.; El-assady, M. QuestionComb: A Gamification Approach for the Visual Explanation of Linguistic Phenomena through Interactive Labeling. *ACM Trans. Interact. Intell. Syst.* **2021**, *11*, 1–38. [CrossRef]
58. Mohseni, S.; Block, J.E.; Ragan, E. Quantitative Evaluation of Machine Learning Explanations: A Human-Grounded Benchmark. In Proceedings of the 26th International Conference on Intelligent User Interfaces, College Station, TX, USA, 14–17 April 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 22–31. [CrossRef]
59. DeYoung, J.; Jain, S.; Rajani, N.; Lehman, E.; Xiong, C.; Socher, R.; Wallace, B. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 4443–4458. [CrossRef]
60. Schuessler, M.; Weiß, P.; Sixt, L. Two4Two: Evaluating Interpretable Machine Learning—A Synthetic Dataset For Controlled Experiments. *arXiv* **2021**, arXiv:2105.02825.
61. Hase, P.; Bansal, M. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 5540–5552. [CrossRef]

62. Nguyen, D. Comparing Automatic and Human Evaluation of Local Explanations for Text Classification. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, LA, USA, 1–6 June 2018; Association for Computational Linguistics: New Orleans, LA, USA, 2018; pp. 1069–1078. [CrossRef]
63. Nauta, M.; Trienes, J.; Pathak, S.; Nguyen, E.; Peters, M.; Schmitt, Y.; Schlötterer, J.; van Keulen, M.; Seifert, C. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *arXiv* **2022**, arXiv:2201.08164.
64. Lu, X.; Tolmachev, A.; Yamamoto, T.; Takeuchi, K.; Okajima, S.; Takebayashi, T.; Maruhashi, K.; Kashima, H. Crowdsourcing Evaluation of Saliency-based XAI Methods. In Proceedings of the ECML PKDD: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Bilbao, Spain, 13–17 September 2021.
65. Fel, T.; Colin, J.; Cadène, R.; Serre, T. What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods. *arXiv* **2021**, arXiv:2112.04417.
66. Schuff, H.; Adel, H.; Vu, N.T. F1 is Not Enough! Models and Evaluation Towards User-Centered Explainable Question Answering. *arXiv* **2020**, arXiv:2010.06283.
67. Friedler, S.A.; Roy, C.D.; Scheidegger, C.; Slack, D. Assessing the Local Interpretability of Machine Learning Models. *arXiv* **2019**, arXiv:1902.03501.
68. Yu, H.; Taube, H.; Evans, J.A.; Varshney, L.R. Human Evaluation of Interpretability: The Case of AI-Generated Music Knowledge. *arXiv* **2020**, arXiv:2004.06894.
69. Heimerl, A.; Weitz, K.; Baur, T.; Andre, E. Unraveling ML Models of Emotion with NOVA: Multi-Level Explainable AI for Non-Experts. *IEEE Trans. Affect. Comput.* 2020, *early access*. [CrossRef]
70. Wang, Y.; Venkatesh, P.; Lim, B.Y. Interpretable Directed Diversity: Leveraging Model Explanations for Iterative Crowd Ideation. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April–5 May 2022.
71. Soltani, S.; Kaufman, R.; Pazzani, M. User-Centric Enhancements to Explainable AI Algorithms for Image Classification. In Proceedings of the Annual Meeting of the Cognitive Science Society, Toronto, ON, Canada, 27–30 July 2022; p. 44.
72. Rebanal, J.C.; Tang, Y.; Combitsis, J.; Chang, K.; Chen, X.A. XAlgo: Explaining the Internal States of Algorithms via Question Answering. *arXiv* **2020**, arXiv:2007.07407.
73. Zhao, W.; Oyama, S.; Kurihara, M. Generating Natural Counterfactual Visual Explanations. In Proceedings of the IJCAI'20: Twenty-Ninth International Joint Conference on Artificial Intelligence, Yokohama, Japan 7–15 January 2021.
74. Ray, A.; Burachas, G.; Yao, Y.; Divakaran, A. Lucid Explanations Help: Using a Human-AI Image-Guessing Game to Evaluate Machine Explanation Helpfulness. *arXiv* **2019**, arXiv:1904.03285.
75. Alvarez-Melis, D.; Kaur, H.; III, H.D.; Wallach, H.M.; Vaughan, J.W. A Human-Centered Interpretability Framework Based on Weight of Evidence. *arXiv* **2021**, arXiv:2104.13299v2
76. Zeng, X.; Song, F.; Li, Z.; Chusap, K.; Liu, C. Human-in-the-Loop Model Explanation via Verbatim Boundary Identification in Generated Neighborhoods. In Proceedings of the Machine Learning and Knowledge Extraction, Virtual Event, 17–20 August 2021; Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 309–327.
77. Baur, T.; Heimerl, A.; Lingenfelser, F.; Wagner, J.; Valstar, M.; Schuller, B.; Andre, E. eXplainable Cooperative Machine Learning with NOVA. *KI Künstliche Intell.* **2020**, *34*, 143–164. [CrossRef]
78. Zöller, M.A.; Titov, W.; Schlegel, T.; Huber, M.F. XAutoML: A Visual Analytics Tool for Establishing Trust in Automated Machine Learning. *arXiv* **2022**, arXiv:2202.11954. [CrossRef]
79. de Bie, K.; Lucic, A.; Haned, H. To Trust or Not to Trust a Regressor: Estimating and Explaining Trustworthiness of Regression Predictions. *arXiv* **2021**, arXiv:2104.06982.
80. Nourani, M.; Roy, C.; Rahman, T.; Ragan, E.D.; Ruozzi, N.; Gogate, V. Don't Explain without Verifying Veracity: An Evaluation of Explainable AI with Video Activity Recognition. *arXiv* **2020**, arXiv:2005.02335.
81. Schmidt, P.; Biessmann, F. Calibrating Human-AI Collaboration: Impact of Risk, Ambiguity and Transparency on Algorithmic Bias. In Proceedings of the Machine Learning and Knowledge Extraction, Dublin, Ireland, 25–28 August 2020; Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 431–449.
82. Bauer, K.; von Zahn, M.; Hinz, O. *Expl(Ai)Ned: The Impact of Explainable Artificial Intelligence on Cognitive Processes*; Working Paper Series; Leibniz Institute for Financial Research SAFE: Frankfurt, Germany, 2021.
83. Jin, W.; Fan, J.; Gromala, D.; Pasquier, P.; Hamarneh, G. EUCA: A Practical Prototyping Framework towards End-User-Centered Explainable Artificial Intelligence. *arXiv* **2021**, arXiv:2102.02437.
84. Chen, C.; Feng, S.; Sharma, A.; Tan, C. Machine Explanations and Human Understanding. *arXiv* **2022**, arXiv:2202.04092. [CrossRef]
85. Zhang, Z.; Singh, J.; Gadiraju, U.; Anand, A. Dissonance Between Human and Machine Understanding. *Proc. ACM Hum.-Comput. Interact.* **2019**, *3*, 1–23. [CrossRef]
86. Anand, A.; Bizer, K.; Erlei, A.; Gadiraju, U.; Heinze, C.; Meub, L.; Nejdl, W.; Steinroetter, B. Effects of Algorithmic Decision-Making and Interpretability on Human Behavior: Experiments using Crowdsourcing. In Proceedings of the 6th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2018), Zurich, Switzerland, 5–8 July 2018.
87. Smith, A.; Nolan, J. The Problem of Explanations without User Feedback. *CEUR Workshop Proc.* **2018**, *2068*, 1–3.

88. Fulton, L.B.; Lee, J.Y.; Wang, Q.; Yuan, Z.; Hammer, J.; Perer, A. Getting Playful with Explainable AI: Games with a Purpose to Improve Human Understanding of AI. In Proceedings of the CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 1–8. [CrossRef]
89. Ghai, B.; Liao, Q.V.; Zhang, Y.; Bellamy, R.K.E.; Mueller, K. Explainable Active Learning (XAL): An Empirical Study of How Local Explanations Impact Annotator Experience. *arXiv* **2020**, arXiv:2001.09219.
90. Linder, R.; Mohseni, S.; Yang, F.; Pentyala, S.K.; Ragan, E.D.; Hu, X.B. How level of explanation detail affects human performance in interpretable intelligent systems: A study on explainable fact checking. *Appl. AI Lett.* **2021**, *2*, e49. [CrossRef]
91. Ehsan, U.; Riedl, M.O. Human-centered Explainable AI: Towards a Reflective Sociotechnical Approach. In Proceedings of the International Conference on Human-Computer Interaction, Copenhagen, Denmark, 19–24 July 2020.
92. Kumar, A.; Vasileiou, S.L.; Bancilhon, M.; Ottley, A.; Yeoh, W. VizXP: A Visualization Framework for Conveying Explanations to Users in Model Reconciliation Problems. *Proc. Int. Conf. Autom. Plan. Sched.* **2022**, *32*, 701–709.
93. Ribera Turró, M.; Lapedriza, A. Can we do better explanations? A proposal of User-Centered Explainable AI. In Proceedings of the ACM IUI 2019 Workshops, Los Angeles, CA, USA, 20 March 2019.
94. Cabour, G.; Morales, A.; Ledoux, E.; Bassetto, S. Towards an Explanation Space to Align Humans and Explainable-AI Teamwork. *arXiv* **2021**, arXiv:2106.01503. [CrossRef]
95. Hohman, F.; Head, A.; Caruana, R.; DeLine, R.; Drucker, S.M. Gamut: A Design Probe to Understand How Data Scientists Understand 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. 1–13. [CrossRef]
96. Nourani, M.; King, J.T.; Ragan, E.D. The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Online, 26–28 October 2020.
97. Chu, E.; Roy, D.; Andreas, J. Are Visual Explanations Useful? A Case Study in Model-in-the-Loop Prediction. *arXiv* **2020**, arXiv:2007.12248.
98. Wang, D.; Zhang, W.; Lim, B.Y. Show or Suppress? Managing Input Uncertainty in Machine Learning Model Explanations. *arXiv* **2021**, arXiv:2101.09498. [CrossRef]
99. Shen, H.; Huang, T.H. How Useful Are the Machine-Generated Interpretations to General Users? A Human Evaluation on Guessing the Incorrectly Predicted Labels. *Proc. AAAI Conf. Hum. Comput. Crowdsourcing* **2020**, *8*, 168–172.
100. Dinu, J.; Bigham, J.P.; Kolter, J.Z. Challenging common interpretability assumptions in feature attribution explanations. *arXiv* **2020**, arXiv:2005.02748
101. Yang, F.; Huang, Z.; Scholtz, J.; Arendt, D.L. How Do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning? In Proceedings of the 25th International Conference on Intelligent User Interfaces, Cagliari, Italy, 17–20 March 2020. [CrossRef]
102. Ferguson, W.; Batra, D.; Mooney, R.; Parikh, D.; Torralba, A.; Bau, D.; Diller, D.; Fasching, J.; Fiotto-Kaufman, J.; Goyal, Y.; et al. Reframing explanation as an interactive medium: The EQUAS (Explainable QUestion Answering System) project. *Appl. AI Lett.* **2021**, *2*, e60, [CrossRef]
103. Coma-Puig, B.; Carmona, J. An Iterative Approach based on Explainability to Improve the Learning of Fraud Detection Models. *CoRR* **2020**, abs/2009.13437. Available online: https://scholar.google.com.sg/scholar?hl=en&as_sdt=0%2C5&q=An+Iterative+Approach+based+on+Explainability+to+Improve+the+Learning+++of+Fraud+Detection+Models&btnG= (accessed on 20 May 2022).
104. Kouvela, M.; Dimitriadis, I.; Vakali, A. Bot-Detective: An Explainable Twitter Bot Detection Service with Crowdsourcing Functionalities. In Proceedings of the 12th International Conference on Management of Digital EcoSystems, Online, 2–4 November 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 55–63. [CrossRef]
105. Collaris, D.; van Wijk, J. ExplainExplore: Visual Exploration of Machine Learning Explanations. In Proceedings of the 2020 IEEE Pacific Visualization Symposium (PacificVis), Tianjin, China, 3–5 June 2020; pp. 26–35. [CrossRef]
106. Yang, Y.; Kandogan, E.; Li, Y.; Sen, P.; Lasecki, W.S. A Study on Interaction in Human-in-the-Loop Machine Learning for Text Analytics. In Proceedings of the IUI Workshops, Los Angeles, CA, USA, 20 March 2019.
107. Arous, I.; Dolamic, L.; Yang, J.; Bhardwaj, A.; Cuccu, G.; Cudré-Mauroux, P. MARTA: Leveraging Human Rationales for Explainable Text Classification. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 5868–5876.
108. Confalonieri, R.; Weyde, T.; Besold, T.R.; Moscoso del Prado Martín, F. Using ontologies to enhance human understandability of global post-hoc explanations of black-box models. *Artif. Intell.* **2021**, *296*, 103471. [CrossRef]
109. El-Assady, M.; Sperrle, F.; Deussen, O.; Keim, D.; Collins, C. Visual Analytics for Topic Model Optimization based on User-Steerable Speculative Execution. *IEEE Trans. Vis. Comput. Graph.* **2019**, *25*, 374–384. [CrossRef]
110. Correia, A.H.C.; Lecue, F. Human-in-the-Loop Feature Selection. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 2438–2445. [CrossRef]
111. Spinner, T.; Schlegel, U.; Schäfer, H.; El-Assady, M. explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. *IEEE Trans. Vis. Comput. Graph.* **2020**, *26*, 1064–1074. [CrossRef]
112. Hadash, S.; Willemsen, M.; Snijders, C.; IJsselsteijn, W. Improving understandability of feature contributions in model-agnostic explainable AI tools. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April–5 May 2022. [CrossRef]

113. Lertvittayakumjorn, P.; Specia, L.; Toni, F. FIND: Human-in-the-Loop Debugging Deep Text Classifiers. *arXiv* **2020**, arXiv:2010.04987.
114. Hohman, F.; Srinivasan, A.; Drucker, S.M. TeleGam: Combining Visualization and Verbalization for Interpretable Machine Learning. In Proceedings of the IEEE Visualization Conference (VIS), Vancouver, BC, Canada, 20–25 October 2019.
115. Guo, L.; Daly, E.M.; Alkan, O.; Mattetti, M.; Cornec, O.; Knijnenburg, B. Building Trust in Interactive Machine Learning via User Contributed Interpretable Rules. In Proceedings of the 27th International Conference on Intelligent User Interfaces, Helsinki, Finland, 22–25 March 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 537–548. [CrossRef]
116. Nushi, B.; Kamar, E.; Horvitz, E. Towards Accountable AI: Hybrid Human-Machine Analyses for Characterizing System Failure. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Zurich, Switzerland, 5–8 July 2018.
117. Liu, Z.; Guo, Y.; Mahmud, J. When and Why does a Model Fail? A Human-in-the-loop Error Detection Framework for Sentiment Analysis. *arXiv* **2021**, arXiv:2106.00954.
118. Balayn, A.; Rikalo, N.; Lofi, C.; Yang, J.; Bozzon, A. How Can Explainability Methods Be Used to Support Bug Identification in Computer Vision Models? In Proceedings of the CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April–5 May 2022; Association for Computing Machinery: New York, NY, USA, 2022. [CrossRef]
119. Afzal, S.; Chaudhary, A.; Gupta, N.; Patel, H.; Spina, C.; Wang, D. Data-Debugging Through Interactive Visual Explanations. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Delhi, India, 11 May 2021; pp. 133–142. [CrossRef]