A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems

SINA MOHSENI and NILOOFAR ZAREI, Texas A&M University
ERIC D. RAGAN, University of Florida

The need for interpretable and accountable intelligent systems grows along with the prevalence of artificial intelligence applications used in everyday life. Explainable AI systems are intended to self-explain the reasoning behind system decisions and predictions. Researchers from different disciplines work together to define, design, and evaluate explainable systems. However, scholars from different disciplines focus on different objectives and fairly independent topics of Explainable AI research, which poses challenges for identifying appropriate design and evaluation methodology and consolidating knowledge across efforts. To this end, this paper presents a survey and framework intended to share knowledge and experiences of Explainable AI design and evaluation methods across multiple disciplines. Aiming to support diverse design goals and evaluation methods in XAI research, after a thorough review of Explainable AI related papers in the fields of machine learning, visualization, and human-computer interaction, we present a categorization of Explainable AI design goals and evaluation methods. Our categorization presents the mapping between design goals for different Explainable AI user groups and their evaluation methods. From our findings, we develop a framework with step-by-step design guidelines paired with evaluation methods to close the iterative design and evaluation cycles in multidisciplinary Explainable AI teams. Further, we provide summarized ready-to-use tables of evaluation methods and recommendations for different goals in Explainable AI research.

Additional Key Words and Phrases: Explainable artificial intelligence (XAI); human-computer interaction (HCI); machine learning; explanation; transparency;

ACM Reference Format:
Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2020. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Trans. Interact. Intell. Syst.* 1, 1, Article 1 (January 2020), 46 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Impressive applications of Artificial Intelligence (AI) and machine learning have become prevalent in our time. Tech giants like Google, Facebook, and Amazon have collected and analyzed enough personal data through smartphones, personal assistant devices, and social media that can model individuals better than other people. Recent negative interference of social media bots in political elections [91, 212] were yet another sign of how susceptible our lives are to the misuse of artificial intelligence and big data [163]. In these circumstances, despite tech giants and the thirst for more advanced systems, others suggest holding off on fully unleashing AI for critical applications until they can be better understood by those who will rely on them. The demand for predictable and

Authors' addresses: Sina Mohseni, sina.mohseni@tamu.edu; Niloofar Zarei, n.zarei.3001@tamu.edu, Texas A&M University, College Station, Texas; Eric D. Ragan, eragan@ufl.edu, University of Florida, Gainesville, Florida.

2160-6455/2020/1-ART1 $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

accountable AI grows as tasks with higher sensitivity and social impact are more commonly entrusted to AI services. Hence, algorithm transparency is an essential factor in holding organizations responsible and accountable for their products, services, and communication of information.

Explainable Artificial Intelligence (XAI) systems are a possible solution towards accountable AI, making it possible by explaining AI decision-making processes and logic for end users [72]. Specifically, explainable algorithms can enable control and oversight in case of adverse or unwanted effects, such as biased decision-making or social discrimination. An XAI system can be defined as a self-explanatory intelligent system that describes the reasoning behind its decisions and predictions. The AI explanations (either on-demand explanations or in the form of model description) could benefit users in many ways such as improving safety and fairness when relying on AI decisions.

While the increasing impact of advanced black-box machine learning systems in the big-data era has attracted much attention from different communities, interpretability of intelligent systems has also been studied in numerous contexts [69, 167]. The study of personalized agents, recommendation systems, and critical decision-making tasks (e.g., medical analysis, powergrid control) has added to the importance of machine-learning explanation and AI transparency for end-users. For instance, as a step towards this goal, the legal right to explanations has been established in the European Union General Data Protection Regulation (GDPR) commission. While the current state of regulations is mainly focused on user data protection and privacy, it is expected to cover more algorithmic transparency and explanations requirements from AI systems [67].

Clearly, addressing such a broad array of definitions and expectations for XAI requires multidisciplinary research efforts, as existing communities have different requirements and often have drastically different priorities and areas of specialization. For instance, research in the domain of machine learning seeks to design new interpretable models and explain black -box models with ad-hoc explainers. Along the same line but with different approaches, researchers in visual analytics design and study tools and methods for data and domain experts to visualize complex black-box models and study interactions to manipulate machine learning models. In contrast, research in human-computer interaction (HCI) focuses on end-user needs such as user trust and understanding of machine generated explanations. Psychology research also studies the fundamentals of human understanding, interpretability, and the structure of explanations.

Looking at the broad spectrum of research on XAI, it is evident that scholars from different disciplines have different goals in mind. Even though different aspects of XAI research are following the general goals of AI interpretability, researchers in each discipline use different measures and metrics to evaluate the XAI goals. For example, numerical analytic methods are employed in machine learning fields to evaluate computational interpretability, while human interpretability and human-subjects evaluations are more commonly the primary goals in HCI and visualization communities. In this regard, although there seems to be a mismatch in specific objectives for designing and evaluating explainability and interpretability, a convergence in goals is beneficial for achieving the full potential of XAI. To this end, this paper presents a survey and framework intended to share knowledge and experiences of XAI design and evaluation methods across multiple disciplines. To support the diverse design goals and evaluation methods in XAI research, after a thorough review of XAI related papers in the fields of machine learning, visualization, and HCI, we present a categorization of interpretable machine learning design goals and evaluation methods and show a mapping between design goals for different XAI user groups and their evaluation methods. From our findings, we develop a framework with step-by-step design guidelines paired with evaluation methods to close the iterative design and evaluation loops in multidisciplinary teams. Further, we provide summarized ready-to-use evaluation methods for different goals in XAI research. Lastly, we review recommendations for XAI design and evaluation drawn from our literature review.

2 BACKGROUND

Nowadays, algorithms analyze user data and affect decision-making processes for millions of people on matters like employment, insurance rates, loan rates, and even criminal justice [35]. However, these algorithms that serve critical roles in many industries have their own disadvantages that can result in discrimination [44, 196], and unfair decision-making [163]. For instance, recently, news feed and targeted advertising algorithms in social media have attracted much attention for aggravating the lack of information diversity in social media [23]. A significant part of the trouble could be because algorithmic decision-making systems—unlike recommender systems—do not allow their users to choose between the recommended items, but instead, present the most relevant content or option themselves. To address this, Heer [75] suggests the use of shared representations of tasks that are augmented with both machine learning models and user knowledge to reduce negative effects of immature AI autonomous systems. They present case studies of interactive systems that integrate proactive computational support into interactive systems.

Bellotti and Edwards [16] argue that intelligent context-aware systems should not act on our behalf. They suggest user control over the system as a principle to support the accountability of a system and its users. Transparency can provide essential information for decision-making that is hidden to the end-users and prevents blind faith [218]. The key benefits of algorithmic transparency and interpretability include: user awareness [9]; bias and discrimination detection [45, 196]; interpretable behavior of intelligent systems [124]; and accountability for users [46]. Furthermore, considering the growing body of examples of discrimination and other legal aspects of algorithmic decision making, researchers are demanding and investigating transparency and accountability of AI under the law to mitigate adverse effects of algorithmic decision making [49, 145, 201]. In this section, we review research background related to XAI systems from a broad and multidisciplinary perspective. At the end, we relate the summaries and positions derived through our survey to other work in the field.

2.1 Auditing Inexplicable AI

Researchers audit algorithms to study bias and discrimination in algorithmic decision making [184] and study the users' awareness of the effects of these algorithms [58]. *Auditing* of algorithms is a mechanism for investigating algorithms' functionality to detect bias and other unwanted algorithm behaviors without the need to know about its specific design details. Auditing methods focus on problematic effects on the results of algorithmic decision-making systems. To audit an algorithm, researchers feed new inputs to the algorithm and review system output and behavior. Researchers generate new data and user accounts with the help of scripts, bots [44], and crowdsourcing [73] to emulate real data and real users in the auditing process. For bias detection among multiple algorithms, cross-platform auditing can detect if an algorithm behaves differently from another algorithm. A recent example of cross-platform auditing is a work by Eslami et al. [59], in which they analyzed user reviews in three hotel booking websites to study user awareness of bias in online rating algorithms. These examples demonstrate that auditing is a valuable yet time-intensive process that could not be scaled easily to large numbers of algorithms. This calls for new research for more effective solutions toward algorithmic transparency.

2.2 Explainable AI

Along with the methods mentioned above for supporting transparency, machine learning explanations have also become a common approach to achieve transparency in many applications such as social media, e-commerce, and data-driven management of human workers [116, 197, 199]. The XAI system, as illustrated in Figure 1, is able to generate explanations and describe the reasoning

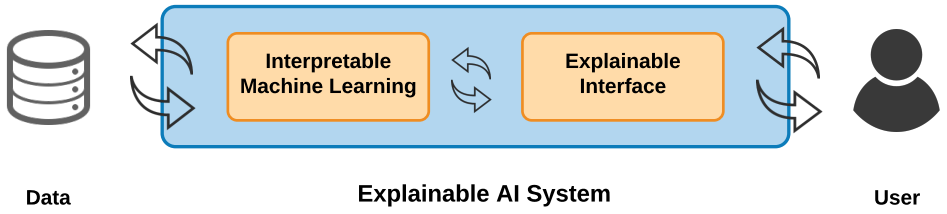

Fig. 1. The user interacts with the explainable interface to send queries to the interpretable machine learning and receive model prediction and explanations. The interpretable model interacts with the data to generate explanation or new new prediction for the user query.

behind machine-learning decisions and predictions. Machine-learning explanations enable users to understand how the data is processed. They aim to bring awareness to possible bias and system malfunctions. For example, to measure user perception of justice in intelligent decision making, Binns et al. [20] studied explanations in systems for everyday tasks such as determining car insurance rates and loan application approvals. Their results highlight the importance of machine learning explanations in users' comprehension and trust in algorithmic decision-making systems. In a similar work studying knowledge of social media algorithms, Radar et al. [170] ran a crowdsourced study to see how different types of explanations affect users' beliefs on news feed algorithmic transparency in a social media platform. In their study, they measured users' awareness, correctness, and accountability to evaluate algorithmic transparency. They found that all explanations caused users to become more aware of the system's behavior. Stumpf et al. [194] designed experiments to investigate meaningful explanations and interactions to hold users accountable by machine learning algorithms. They show explanations as a potential method for supporting richer human-computer collaboration to share intelligence.

The recent advancements and trends for explainable AI research demand a wide range of goals for algorithmic transparency which calls for research across varied application areas. To this end, our review encourages a cross-discipline perspective of intelligibility and transparency goals.

2.3 Related Surveys and Guidelines

In recent years, there have been surveys and position papers suggesting research directions and highlighting challenges in interpretable machine learning research [48, 78, 127]. Although our review is limited to computer science literature, here we summarize several of the most relevant peer-reviewed surveys related to the topic of XAI across active disciplines including social science. While all surveys, models, and guidelines in this section add value to the XAI research, to the best of our knowledge, there is no existing comprehensive survey and framework for evaluation methods of explainable machine learning systems.

2.3.1 Social Science Surveys. Research in the social sciences is particularly important for XAI systems to understand how people generate, communicate, and understand explanations by taking into account each others' thinking, cognitive biases, and social expectations in the process of explaining. Hoffman, Mueller, and Klein reviewed the key concepts of explanations for intelligent systems in a series of essays to identify how people formulate and accept explanations, ways to generate self-explanations, and identified purposes and patterns for causal reasoning [83, 84, 102]. They lastly focus on deep neural networks (DNN) to examine their theoretical and empirical findings on a machine learning algorithm [79]. In other work, they presented a conceptual model of the process of explaining in the XAI context [85]. Their framework includes specific steps and

measures for the goodness of explanations, user satisfaction and understanding of explanations, users' trust and reliance on XAI systems, effects of curiosity on the search for explanations, and human-XAI system performance.

Miller [142] suggests a close collaboration between machine learning researchers in the space of XAI with social science would further refine the explainability of AI for people. He discusses how understanding and replicating how people generate, select, and present explanations could improve human-XAI interactions. For instance, Miller reviews how people generate and select explanations that are involved with cognitive biases and social expectations. Other papers reviewing social science aspects of XAI systems include studies on the role of algorithmic transparency and explanation in lawful AI [49] and of fair and accountable algorithmic decision-making processes [117].

2.3.2 Human Computer Interactions Surveys. Many HCI surveys discuss the limitations and challenges in AI transparency [208] and interactive machine learning [6]. Others suggest a set of theoretical and design principles to support intelligibility of intelligent system and accountability of human users (e.g., [16, 90]). In a recent survey, Abdul et al. [1] presented a thorough literature analysis to find XAI-related topics and relationships among these topics. They used visualization of keywords, topic models, and citation networks to present a holistic view of research efforts in a wide range of XAI related domains; from privacy and fairness to intelligent agents and context-aware systems. In another work, Wang et al. [204] explored theoretical underpinnings of human decision-making and proposed a conceptual framework for building human-centered decision-theory-driven XAI systems. Their framework helps to choose better explanations to present, backed by reasoning theories, and human cognitive biases. Focused on XAI interface design, Eiband et al. [56] present a stage-based participatory process for integration of transparency in existing intelligent systems using explanations. Another design framework is XAID from Zhu et al. [225], which presents a human-centered approach for facilitating game designers to co-create with machine learning techniques. Their study investigates the usability of XAI algorithms in terms of how well they support game designers.

2.3.3 Visual Analytics Surveys. XAI-related surveys in the visualization domain follow visual analytics goals such as understanding and interacting with machine learning systems in different visual analytics applications [57, 180]. Choo and Liu [34] reviewed challenges and opportunities for Visual Analytics for explainable deep learning design. In a recent paper, Hohman et al. [88] provide an excellent review and categorization of visual analytics tools for deep learning applications. They cover various data and visualization techniques that are being used in deep visual analytics applications. Also, Spinner et al. [192] proposed a XAI pipeline which maps the XAI process to an iterative workflow in three stages: model understanding, diagnosis, and refinement. To operationalize their framework, they designed explAIner, a visual analytics system for interactive and interpretable machine learning that instantiates all steps of their pipeline.

2.3.4 Machine Learning Surveys. In the machine learning area, Guidotti et al. [71] present a comprehensive review and classification of machine learning interpretability techniques. Also, Montavon et al. [152] focus on interpretability techniques for DNN models. On Convolutional Neural Network (CNN), Zhang et al. [221] reviews research on interpretability techniques in six directions including visualization of CNN representations, diagnosing techniques for CNNs, approaches for transforming CNN representations into interpretable graphs, building explainable models, and semantic-level learning based on model interpretability. In another work, Gilpin et al. [64] reviews interpretability techniques in machine learning algorithms and categorizes evaluation approaches to bridge the gap between machine learning and HCI communities.

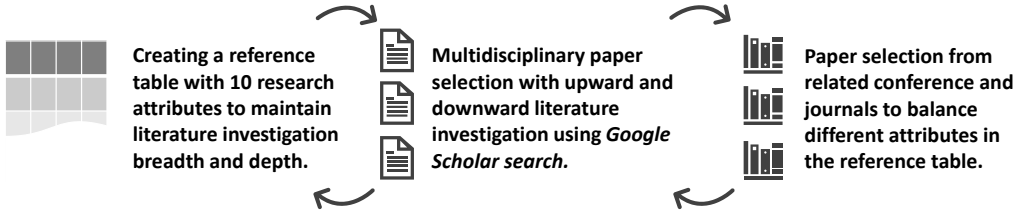

Fig. 2. A diagram summarizing our iterative and multi-pass literature selection and review process to achieve desired literature investigation breadth and depth. We started with 40 papers to create the reference table. Then we added 80 papers by upward and downward literature investigation to improve review breath and depth. Finally, we added another 80 papers from related conferences proceedings and journals to balance the reference table.

In complementing the existing work, our survey provides a multidisciplinary categorization of design goals and evaluation methods for XAI systems. As a result of surveyed papers, we propose a framework that provides a step-by-step design and evaluation plan for a multidisciplinary team of designers for building real-world XAI systems. Unlike Eiband et al. [56], we do not make the assumption of *adding transparency* to *an existing* intelligent interface and do not limit the evaluation of XAI systems to the users' mental model. We instead characterize both design goals and evaluation methods and compile all in a unified framework for multidisciplinary teamwork. Our design framework has similarities to Wang et al's [204] theoretical framework which supports our design goals (see Section 9.6). Our multidisciplinary work extends their conceptual framework by 1) including the design of interpretability algorithms as part of the framework and 2) pairing evaluation methods with each design step to close the iterative design and evaluation loops.

3 SURVEY METHOD

We conducted a survey of the existing research literature to capture and organize the breadth of designs and goals for XAI evaluation. We used a structured and iterative methodology to find XAI-relevant research and categorize the evaluation methods presented in research articles (summarized in Figure 2). In our iterative paper selection process, we started by selecting existing work from top computer science conferences and journals across the fields of HCI, visualization, and machine learning. However, since XAI is a quite fast growing topic, we also wanted to include *arXiv* preprints and useful discussions in workshop papers. We started with 40 papers related to XAI topics across three research fields including but not limited to research on interpretable machine learning techniques, deep learning visualization, interactive model visualization, machine explanations in intelligent agents and context-aware systems, explainable user interfaces, explanatory debugging, and algorithmic transparency and fairness.

We then used selective coding to identify 10 main research attributes in those papers. The main attributes we identified include: research discipline (social science, HCI, visualization, or machine learning), paper type (interface design, algorithm design, or evaluation paper), application domain (machine learning interpretability, algorithmic fairness, recommendation systems, transparency of intelligent systems, intelligent interactive systems and agents, explainable intelligent systems and agents, human explanations, or human trust), machine learning model (e.g., deep learning, decision trees, SVM), data modality (image, text, tabular data), explanation type (e.g., graphical, textual, data visualization), design goal (e.g., model debugging, user reliance, bias mitigation), evaluation type (e.g., qualitative, computational, quantitative with human-subjects), targeted user (AI novices, data experts, AI experts), and evaluation measure (e.g., user trust, task performance, user mental model).

In the second round of collecting XAI literature, we conducted an upward and downward literature investigation using the *Google Scholar* search engine to add 80 more papers to our reference table. We narrowed down our search by XAI related topics and keywords including but not limited to: interpretability, explainability, intelligibility, transparency, algorithmic decision-making, fairness, trust, mental model, and debugging in machine learning and intelligent systems. With this information, we performed axial coding to organize the literature and started discussions on our proposed design and evaluation categorization.

Finally, to maintain reasonable literature coverage and balance the number of papers for each of our categories of design goals and evaluation measures, we added another 80 papers to our reference table. The conferences from which we selected XAI related paper were including but not limited to: CHI, IUI, HCOMP, SIGDIAL, UbiComp, AIES, VIS, ICWSM, IJCAI, KDD, AAAI, CVPR, and NeurIPS conferences. The journals included: Trends in cognitive science, Transactions on Cognitive and Developmental Systems, Cognition Journal, Transactions on Interactive Intelligent Systems, International Journal of Human-Computer Studies, Transactions on Visualization and Computer Graphics, and Transactions on Neural Networks and Learning Systems.

Following a review of 226 papers, our categorization of XAI design goals and evaluation methods is supported by references from papers performing design or evaluation of XAI systems. Our reference table[1] is available online to the research community to provide further insight beyond our discussions in this document. Table 2 shows a digest of our surveyed papers that contains 42 papers with both design and evaluation of XAI system. Later in the Section 7, we provide a series of tables to organize different evaluation methods from research papers with example references for each, documenting our in-depth analysis of 69 papers in total.

4 XAI TERMINOLOGY

To familiarize the readers with common XAI concepts and terminologies that are repeatedly referenced in this review, the following four subsections summarize high-level characterizations of model explanations. Many related surveys (e.g., [2, 207]) and reports (e.g., [38, 200]) also provide useful compilations of terminology and concepts in comprehensive reports. For instance, Abdul et al. [1] present a citation graph from diverse domains related to explanations, including intelligible intelligent systems, context-aware systems, and software learnability. Later, Arrieta et al. [11] present a thorough review of XAI concepts and taxonomies and arrives at the concept of *Responsible AI* as a manifold of multiple AI principles including model fairness, explainability, and privacy. Similarly, the concept of *Safe AI* has been reviewed by Amodei et al. [8], which is an interest in safety-critical intelligent applications such as autonomous vehicles [147]. Table 1 presents descriptions for 14 common terms related to this survey's topic and organizes their relation to *Intelligible Systems* and *Transparent AI* topics. We consider Transparent AI systems as the AI-based class of Intelligible Systems. Therefore, properties and goals previously established for Intelligible Systems are ideally transferable to Transparent AI systems. However, challenges and limitations for achieving transparency in complex machine learning algorithms raise issues (e.g., ensuring the fairness of an algorithm) that were not necessarily problematic in intelligible rule-based systems but now require closer attention from research communities.

The descriptions presented in Table 1 are meant to be an introduction to these terms, though exact definitions and interpretations can depend on usage context and research discipline. Consequently, researchers from different disciplines often use these terms interchangeably, disregarding differences in meaning [2]. Perhaps the two generic terms of *Black-box Model* and *Transparent Model* are in the center of XAI terminology ambiguity. The black-box term refers to complex machine learning

[1]https://github.com/SinaMohseni/Awesome-XAI-Evaluation

Table 1. Table of common terminology related to Intelligible Systems and Transparent AI. Higher-level main concepts are shown in gray, while related terms for the main concepts are listed below and categorized as a desired outcome, property, or practical approach. Explainable AI is one particular practical approach for intelligible systems to enable improve transparency. Note that definitions and interpretations can vary across the literature, and this table is meant to serve as a quick reference.

Concept	Category	Description
Intelligible System	Main Concept	A system that is understandable and predictable for its users through transparency or explanations [1, 16, 207].
Understandability (Intelligibility)	Desired Properties	Intelligible systems support user understanding of system's underlying functions [11, 123].
Predictability		Intelligibility supports building a mental model of the system that enables user to predict system behavior [207].
Trustworthiness	Desired Outcomes	Enabling positive user attitude toward the system that emerges from knowledge, experience, and emotion [82, 85].
Reliability		Supporting user trust to rely and follow system's advice for higher performance [82, 85].
Safety		Improving safety by reducing user unintended misuse due to misperception and unawareness [147].
Transparent AI	Main Concept	An AI-based system that provides information about its decision-making processes [38, 127].
Interpretable AI	Practical Approaches	Inherently human-interpretable models due to their low complexity of machine learning algorithms [151].
Explainable AI		Supporting user understanding of complex models by providing explanations for predictions [204].
Interpretability	Desired Properties	The ability to support user understanding and comprehension of the model decision making process and predictions [11, 127].
Explainability		The ability to explain underlying model and its reasoning with accurate and user comprehensible explanations [11, 127].
Accountable AI	Desired Outcomes	Allowing for auditing and documentation to hold organizations accountable for their AI-based products and services [49, 117].
Fair AI		Enabling ethical and fairness analysis of models and data used in decision-making processes [11, 117].

models that are not human-interpretable [127] as opposed to transparent models which are simple enough to be human-interpretable [11]. We find it more accurate and consistent to separate the transparency of an XAI system (as described in Figure 1) from the interpretability of its internal machine learning models. Specifically, Table 1 shows that Transparent AI could be achieved by either *Interpretable AI* or *Explainable AI* approaches. Other examples of terminology ambiguity include the terms *Interpretability* and *Explainability* that are often used as synonyms in the field of machine learning. For example the phrase "interpretable machine learning technique" often refers to ad-hoc techniques for generating explanations for non-interpretable models such as DNNs [151]. Another example is the occasional case of using the terms *Transparent System* and *Explainable System* interchangeably in HCI research [56], while others clarify that explainability is not equivalent to transparency because it does not require knowing the flow of the bits in the AI decision-making process [49].

4.1 Global and Local Explanations

One way to classify explanations is by their interpretation scale. For instance, an explanation could be as thorough as describing the entire machine learning model. Alternatively, it could only

partially explain the model, or it could be limited to explaining an individual input instance. *Global Explanation* (or *Model Explanation*) is an explanation type that describes how the overall machine learning model works. Model visualization [130, 131] and decision rules [113] are examples of explanations falling in this category. In other cases, interpretable approximations of complex models serve as the model explanation. Tree regularization [213] is a recent example of regularized complex model to learn tree-like decision boundaries. Model complexity and explanation design are the main factors used to choose between different types of global explanations.

In contrast, *Local Explanations* (or *Instance Explanations*) aim to explain the relationship between specific input-output pairs or the reasoning behind the results for an individual user query. This type of explanation is thought to be less overwhelming for novices, and it can be suited for investigating edge cases for the model or debugging data. Local explanations often make use of saliency methods [14, 219] or local approximation of the main model [172, 173]. Saliency methods (also as known as attribution maps or sensitivity maps) use different approaches (e.g., perturbation-based methods, gradient-based methods) to show what features in the input strongly influence the model's prediction. Local approximation of the model, on the other hand, trains an interpretable model (learned from the main model) to locally represent the complex model's behavior.

4.2 Interpretable Models vs. Ad-hoc Explainers

The human interpretability of a machine learning model is inversely proportional to the model's size and complexity. Complex models (e.g., deep neural networks) with high performance and robustness in real-world applications are not interpretable by human users due to their large variable space. Linear regression models or decision trees offer better interpretability but have limited performance on high-dimensional data, whereas a random forest model (ensemble of hundreds of decision trees) can have much higher performance but is less understandable. This trade-off between model interpretability and performance led researchers to design ad-hoc methods to explain any black-box machine learning algorithm such as deep neural networks. Ad-hoc explainers (e.g., [134, 172]) are independent algorithms that can describe model predictions by explaining "why" a certain decision has been made instead of describing the whole model. However, there are limitations in explaining black-box models with ad-hoc explainers, such as the uncertainty of the fidelity of the explainer itself. We will discuss more about the fidelity of explanations in Section 7.5. Furthermore, although ad-hoc explainers generally describe "why" a prediction is made, these methods lack in explaining "how" the decision is made.

4.3 What to Explain

When users face a complex intelligent system, they may demand different types of explanatory information and each explanation type may require its own design. Here we review six common types of explanations used in XAI system designs.

How Explanations demonstrate a holistic representation of the machine learning algorithm to explain *How* the model works. For visual representations, model graphs [113] and decision boundaries [135] are common design examples for *How* explanations. However, research shows users may also be able to develop a mental model of the algorithm based on a collection of explanations from multiple individual instances [133].

Why Explanations describe *Why* a prediction is made for a particular input. Such explanations aim to communicate what features in the input data [172] or what logic in the model [113, 173] has led to a given prediction by the algorithm. This type of explanation can have either model agnostic [134, 172] or model dependent [188] solutions.

Why-not Explanations help users to understand the reasons why a specific output was not in the output of the system [202]. *Why-not* explanations (also called *Contrastive Explanations*) characterize the reasons for differences between a model prediction and the user's expected outcome. Feature importance (or feature attribution) is commonly used as an interpretability technique for *Why* and *Why-not* explanations.

What-If Explanations involve demonstration of how different algorithmic and data changes affect model output given new inputs [29], manipulation of inputs [125], or changing model parameters [103]. Different what-if scenarios may be automatically recommended by the system or can be chosen for exploration through interactive user control. Domains with high-dimensional data (e.g., image and text) and complex machine learning models (e.g., DNNs) have fewer parameters for users to directly tune and examine trained model compared to simpler data (e.g., low-dimensional tabular data) and models.

How-to Explanations spell out hypothetical adjustments to the input or model that would result in a different output [125, 126], such as a user-specified output of interest. Techniques to generate *How-to* (or *Counterfactual*) explanations are ad-hoc and model-agnostic considering that model structure and internal values are not a part of the explanation [203]. Such methods can work interactively with the user's curiosity and partial conception of the system to allow an evolving mental model of the system through iterative testing.

What-else Explanations present users with similar instances of input that generate the same or similar outputs from the model. Also called *Explanation by Example*, *What-else* explanations pick samples from the model's training dataset that are similar to the original input in the model representation space [30]. Although very popular and easy to achieve, research shows example-based explanations could be misleading when training datasets lack uniform distribution of the data [98].

4.4 How to Explain

In all types of machine learning explanations, the goal is to reveal new information about the underlying system. In this survey, we mainly focus on human-understandable explanations, though we note that research on interpretable machine learning has also studied other purposes such as knowledge transfer, object localization, and error detection [61, 162].

Explanations can be designed using a variety of formats for different user groups [216]. *Visual Explanations* use visual elements to describe the reasoning behind the machine learning models. Attention maps and visual saliency in the form of saliency heatmaps [190, 219] are examples of visual explanations that are widely used in machine learning literature. Verbal Explanations describe the machine's model or reasoning with words, phrases, or natural language. Verbal explanations are popular in applications like question-answering explanations and decision lists [113]. This form of explanation has also been implemented in recommendation systems [17, 77] and robotics [176]. Explainable interfaces commonly make use of multiple modalities (e.g., visual, verbal, and numerical elements) for explanations to support user understanding [156]. *Analytic Explanation* is another approach to view and explore the data and the machine learning models representations [88]. Analytic explanations commonly rely on numerical metrics and data visualizations. Visual analytics tools also allow researchers to review model structures, relations, and their parameters in complex deep models. Heatmap visualizations [193], graphs and networks [66], and hierarchical (decision trees) visualizations are commonly used to visualize analytic explanations for interpretable algorithms. Recently, Hohman et al. [87] implemented a combination of visualization and verbalization to communicate or summarize key aspects of a model.

From a different perspective, Chromik et al. [36] extends the idea of "dark patterns" from interactive user interface design [68] into machine learning explanations. They review possible

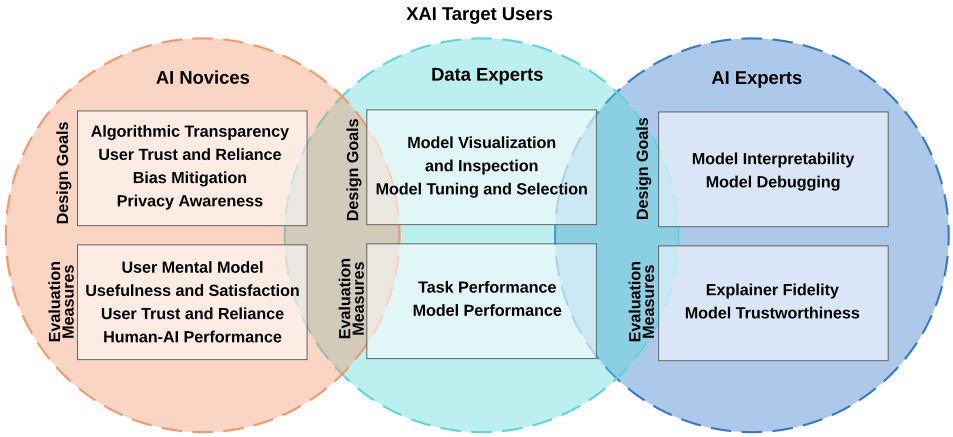

Fig. 3. A summary of our categorization of XAI design goals and evaluation measures between user groups. **Top:** Different system design goals for each user group. **Bottom:** Common evaluation measures used in each user group. Notice that similar XAI goals for different user groups require different research objectives, design methods, and implementation paths.

ways that phrasing of explanations and their implementation in the interface could deceive users for the benefit of other parties. They review negative effects such as lack of user attention to explanations, formation of an incorrect mental model, and even algorithmic anxiety [93] could be among the consequences of such deceptive presentations and interactions of machine learning explanations.

5 CATEGORIZATION OF XAI DESIGN GOALS AND EVALUATION METHODS

While an ideal XAI system should be able to answer all user queries and meet all XAI concept goals [72], individual research efforts focus on designing and studying XAI systems with respect to specific interpretability goals and specific users. Similarly, evaluating the explanations can demonstrate and verify the effectiveness of the explainable systems for their intended goals.

After careful review and analysis of XAI goals and their evaluation methods in the literature, we recognized the following two attributes to be most significant for our purposes of interdisciplinary organization of XAI design and evaluation methods:

- **Design Goals**. The first attribute in our categorization is the design goal for interpretable algorithms and explainable interfaces in XAI research. We obtain XAI design goals from multiple research disciplines: machine learning, data visualization, and HCI. To better understand the differences between various goals for XAI, we organize XAI design goals with their three user groups: AI novices (i.e., general AI product end-user), data experts (experts in data analytics and domain experts), and AI experts (machine learning model designers).
- **Evaluation Measures**. We review evaluation methods and discuss measures used to evaluate machine learning explanations. The measures include user mental model, user trust and reliance, explanation usefulness and satisfaction, human-machine task performance, and computational measures. In our review, we will pay more attention to evaluation measures of XAI as the authors believe that this category is relatively less explored.

Figure 3 presents the pairing between XAI design goals and their evaluation measures. Note that user groups is used as an auxiliary dimension to emphasize on the importance of end users for

system goals. The overlap between XAI user groups shows similarities in the design and evaluation methods between different targeted user groups. However, the similar XAI goals in different user groups require different research objectives, design methods, and implementation paths. To help summarize our characterization along with example literature, Table 2 presents a cross-reference table of XAI evaluation literature to emphasize the importance of design goals, evaluation measures, and user types. We first review details of research focusing on XAI design goals in Section 6 including eight goals organized by their user groups. We then review evaluation measures and methods in Section 7 including six main measures and their methods collected from the surveyed literature.

6 XAI DESIGN GOALS

Research efforts have explored many goals for XAI systems. Doshi-Velez and Kim [48] reviewed multiple high-level priorities for XAI systems with examples including safety, ethics, user reliance, and scientific understanding. Later, Arrieta et al. [11] presented a thorough review of XAI opportunities in different application domains. Accordingly, different design choices such as explanation type, scope, and level of detail will be affected by the application domain, design goal, and user type. For example, while machine learning experts might prefer highly-detailed visualizations of deep models to help them optimize and diagnose algorithms, end-users of daily-used AI products do not expect fully detailed explanations for every query from a personalized agent. Therefore, XAI systems are expected to provide the right type of explanations for the right group of users, meaning it will be most efficient to design an XAI system according to the user's needs and levels of expertise.

To this end, we distinguish XAI design goals based on the designated end-user and evaluation subjects, which we categorize into three general groups of AI experts, data experts, and AI novices. We emphasize that this separation of groups is presented primarily for organizational convenience, as goals are not mutually exclusive across groups, and specific priorities are case dependent for any particular project. The XAI design goals also extend to the broader goal of *Responsible AI* by improving transparency and explainability of intelligent systems. Note that although there are overlaps in the methods used to achieve these goals, the research objectives and design approaches are substantially different among distinct research fields and their user groups. For instance, even though leveraging interpretable models to reduce machine learning model bias is a research goal for AI experts, bias mitigation is also a design goal for AI novices to avoid adverse effects of algorithmic decision-making in their respective domain settings. However, interpretability techniques for AI experts and bias mitigation tools for AI novice require different design methods and elements. In the following subsections, we review eight design goals for XAI systems organized by their user groups.

6.1 AI Novices

AI novices refer to end-users who use AI products in daily life but have no (or very little) expertise on machine learning systems. These include end-users of intelligent applications like personalized agents (e.g., home assistant devices), social media, and e-commerce websites. In most smart systems, machine learning algorithms serve as internal functions and APIs to enable specific features embedded in intelligent and context-aware interfaces. Previous research shows intuitive interface and interaction design can enhance users' experience with the system through improving end-users' comprehension and reliance on the intelligent systems [154]. In this regard, creating human-understandable and yet accurate representations of complicated machine learning explanations for novice end-users is a challenging design trade-off in XAI systems. Note that although there are

Table 2. Tabular summary of our XAI evaluation dimensions of measures and targeted user types. The table includes 42 papers that represent a subset of the surveyed literature organized by the two dimensions.

Work	Design Goals: Novice Users				Design Goals: Data Experts		Design Goals: AI Experts		Evaluation Measures				
	G1: Algorithmic Transparency	G2: User Trust and Reliance	G3: Bias Mitigation	G4: Privacy Awareness	G5: Model Visualization and Inspection	G6: Model Tuning and Selection	G7: Model Interpretability	G8: Model Debugging	M1: Mental Model	M2: Usefulness and Satisfaction	M3: User Trust and Reliance	M4: Human-AI Task Performance	M5: Computational Measures
Herlocker et al. 2000 [77]		◆								◆	◆	◆	
Kulesza et al. 2012 [109]	◆								◆	◆		◆	
Lim et al. 2009 [124]	◆								◆	◆			
Stumpf et al. 2018 [195]	◆	◆							◆		◆		
Bilgic et al. 2005[19]		◆								◆	◆		
Bunt et al. 2012 [25]	◆									◆			
Gedikli et al. 2014 [62]		◆								◆			
Kulesza et al. 2013 [111]	◆	◆							◆		◆		
Lim et al. 2009 [125]	◆	◆							◆	◆	◆	◆	
Lage et al. 2019 [112]	◆									◆		◆	
Schmid et al. 2016 [186]	◆											◆	
Berkovsky et al. 2017 [17]		◆								◆	◆		
Glass et al. 2008 [65]		◆								◆	◆		
Haynes et al. 2009 [74]		◆								◆	◆		
Holliday et al. 2016 [89]	◆	◆							◆		◆		
Nothdurft et al. 2014 [158]	◆	◆							◆		◆		
Pu and Chen et al. 2006 [169]		◆									◆	◆	
Bussone et al. 2015 [26]	◆	◆									◆		
Groce et al. 2014 [70]	◆								◆			◆	
Myers et al. 2006 [156]	◆								◆			◆	
Binns et al. 2018 [20]	◆		◆						◆				
Lee et al. 2019 [115]	◆		◆						◆	◆			
Rader et al. 2018 [170]	◆			◆					◆	◆			
Datta et al. 2015 [44]				◆									◆
Kulesza et al. 2015 [108]	◆				◆	◆			◆			◆	
Kulesza et al. 2010 [110]	◆				◆	◆			◆			◆	
Krause et al. 2016 [107]					◆	◆						◆	
Krause et al. 2017 [105]					◆	◆						◆	
Liu et al. 2014 [131]					◆							◆	
Ribeiro et al. 2016 [172]							◆		◆		◆	◆	◆
Ribeiro et al. 2018 [173]							◆		◆		◆	◆	◆
Ross et al. 2017 [177]							◆						◆
Adebayo et al. 2018 [3]							◆						◆
Samek et al. 2017 [183]							◆						◆
Zeiler et al. 2014 [219]							◆					◆	◆
Lakkaraju et al. 2016 [113]							◆					◆	
Kahng et al. 2018 [95]								◆		◆		◆	
Liu et al. 2018 [129]								◆		◆		◆	
Liu 2017 et al. 2009 [130]								◆		◆		◆	
Ming et al. 2017 [143]								◆		◆		◆	
Pezzotti et al. 2018 [165]								◆		◆		◆	
Strobelt et al. 2018 [193]								◆		◆		◆	

overlaps among goals for *AI Novices* and AI experts who build interpretable algorithms, each user group requires a different set of design methods and objectives that are being studied in different research communities.

The main design goals for AI novice end-users of XAI system can be itemized as the following:

G1: Algorithmic Transparency: An immediate goal for a XAI system – in comparison to an inexplicable intelligent system – is to help end-users understand how the intelligent system works. Machine learning explanations improve users' mental model of the underlying intelligent algorithms by providing comprehensible transparency for the complex intelligent algorithms [208]. Further, transparency of a XAI system can improve user experience through better user understanding of model output [123], thus improving user interactions with the system [108].

G2: User Trust and Reliance: XAI system can improve end-users trust in the intelligent algorithm by providing explanations. A XAI system lets users assess system reliability and calibrate their perception of system accuracy. As a result, users' trust in the algorithm leads to their reliance on the system. Example applications where XAI aims to improve user reliance through its transparent design include recommendation systems [17], autonomous systems [209], and critical decision making systems [26] .

G3: Bias Mitigation: Unfair and biased algorithmic decision-making is a critical side effect of intelligent systems. Bias in machine learning has many sources, including biased training data and feature learning that could result in discrimination in algorithmic decision-making [137]. Machine learning explanations can help end-users to inspect if the intelligent systems are biased in their decision-making. Examples of cases in which XAI is used for bias mitigation and fairness assessment are criminal risk assessment [20, 115] and loan and insurance rate prediction [32]. It is worth mentioning that there is overlap between the biased decision-making mitigation goal for AI novices and the goal of dataset bias for AI experts (Section 6.2), which results in shared implementation techniques. However, the two distinct user groups require their own sets of XAI design goals and processes.

G4: Privacy Awareness: Another goal in designing XAI systems is to provide a means for end-users to assess their data privacy. Machine learning explanations can disclose to end-users what user data is being used in algorithmic decision-making. Examples of AI application examples in which privacy awareness is primarily important include personalized advertisements using users' online advertisement [44] and personalized news feed in social media [58, 170].

In addition to the major XAI goals, interactive visualization tools have also been developed to help AI novices to learn machine learning concepts and models by interacting with simplified data and model representations. Examples of these educative tools include TensorFlow PlayGround [191] for teaching elementary neural networks concepts and Adversarial Playground [157] for learning concept of adversarial examples in DNNs. These minor goals cover XAI system objectives that have limited extent compared to main goals.

6.2 Data Experts

Data experts include data scientists and domain experts who use machine learning for analysis, decision-making, or research. Visual analysis tools can support interpretable machine learning in many ways, such as visualizing the network architecture of a trained model and training process

of machine learning models. Researchers have implemented various visualization designs and interaction techniques to understand better and improve machine learning models.

Data experts analyze data in specialized forms and domains, such as cybersecurity [18, 66], medicine [31, 107], text [128, 131], and satellite image analysis [174]. These users might be experts of certain domain areas or experts in general areas of data science, but in our categorization, we consider users in the *data experts* category to generally lack expertise in the technical specifics of the machine learning algorithms. Instead, this group of users often utilize intelligent data analysis tools or visual analytics systems to obtain insights from the data. Notice that there are overlaps between XAI goals in different disciplines and visual analytics tools designed for *Data Experts* could be used by both model designers and data analysts. However, design needs and approaches for these XAI systems may be different across research communities. The main design goals for data experts users of a XAI system are as follows:

G5: Model Visualization and Inspection: Similar to AI novices, data experts also benefit from machine learning interpretability to inspect model uncertainty and trustworthiness [181]. For instance, machine-learning explanations help data experts to visualize models [86] and inspect for problems like bias [4]. Another important aspect of model visualization and inspection for domain experts is to identify and analyze failure cases of machine learning models and systems [144]. Therefore, the main challenge for data-analysis and decision-support systems is to improve model transparency via visualization and interaction techniques for domain experts [216].

G6: Model Tuning and Selection: Visual analytics approaches can help data experts to tune machine learning parameters for their specific data in an interactive visual fashion [131]. The interpretability element in XAI visual analytic systems increase data experts' ability to compare multiple models [5] and select the right model for the targeted data. As an example, Du et al. [51] present EventAction, an event sequence recommendation approach that allows the users to interactively select records that share their desired attribute values. In the case of tuning DNN networks, visual analytics tools enhance designers' ability to modify networks [165], improve training [129], and to compare different networks [211].

6.3 AI Experts

In our categorization, *AI experts* are machine learning scientists and engineers who design machine learning algorithms and interpretability techniques for XAI systems. Machine learning interpretability techniques either provide model interpretation or instance explanations. Examples of model interpretation techniques include inherently interpretable models [205], deep model simplification [213], and visualization of model internals [215]. Instance explanations techniques, however, present feature importance for individual instances such as saliency map in image data and attention in textual data [43]. AI engineers also benefit from visualization and visual analytics tools to interactively inspect model internal variables [129] to detect architecture and training flaws or monitor and control the training process [95], which indicates possible overlaps among design goals. We list main design goals for AI Experts into two following items:

G7: Model Interpretability: Model interpretability is often a primary XAI Goal for AI experts. Model interpretability allows getting new insights into how deep models learn patterns from data [162]. In this regard, various interpretability techniques for different domains have been proposed to satisfy the need for explanation [99, 188]. For example, Yosinski et al. [215] created an interactive toolbox to explore CNN's activation layers in real-time that gives an intuition about

Table 3. Evaluation measures and methods used in studying user mental models in XAI systems

Mental Model Measures	Evaluation Methods
User Understanding of Model	Interview ([40]) and Self-explanation ([20, 47, 164])
	Likert-scale Questionnaire ([99, 111, 113, 124, 133, 171])
Model Output Prediction	User Prediction of Model Output ([96, 172, 173])
Model Failure Prediction	User Prediction of Model Failure ([15, 161])

"how the CNN works" to the user.

G8: Model Debugging: AI researchers use interpretability techniques in different ways to improve model architecture and training process. For example, Zeiler and Fergus [219] present a use case in which visualization of filters and feature maps in CNN leads to revising training hyper-parameters and, therefore, model performance improvement. In another work, Ribeiro et al. [172] used model instance explanations and human review of explanations to improve model performance through feature engineering.

Other than main XAI goals for AI experts, machine learning explanations are used for other purposes including detecting dataset bias [220], adversarial attack detection [61], and model failure prediction [146]. Also, visual saliency maps and attention mechanisms have been used as weakly supervised object localization [190], multiple object recognition [12], and knowledge transfer [122] techniques.

7 XAI EVALUATION MEASURES

Evaluation measures for XAI systems is another important factor in the design process of XAI systems. Explanations are designed to answer different interpretability goals, and hence different measures are needed to verify explanation validity for the intended purpose. For example, experimental design with human-subject studies is a common approach to perform evaluations with AI novice end-users. Various controlled in-lab and online crowdsourced studies have been used for XAI evaluation. Also, case studies aim to collect domain expert users' feedback while performing high-level cognitive tasks with analytics tools. By contrast, computational measures are designed to evaluate the accuracy and completeness of explanations from interpretable algorithms.

In this section, we review and categorize the main evaluation measures for XAI systems and algorithms. Table 2 shows a list of five evaluation measures associated with their design goals. Additionally, we provide summarized and ready-to-use XAI evaluation measures and methods extracted from the literature in Tables 3-7.

7.1 M1: Mental Model

Following cognitive psychology theories, a mental model is a representation of how users understands a system. Researchers in HCI study users' mental models to determine their understanding of intelligent systems in various applications. For example, Costanza et al. [40] studied how users understand a smart grid system, and Kay et al. [96] studied how users understand and adapt to uncertainty in machine learning prediction of bus arrival times.

In the context of XAI, explanations help users to create a mental model of *how the AI works*. Machine learning explanation is a way to help the users in building a more accurate mental model. Studying users' mental models of XAI systems can help verify explanation effectiveness

in describing an algorithm's decision-making process. Table 3 summarizes different evaluation methods used to measure users' mental model of machine learning models.

Psychology research in human-AI interactions has also explored structure, types, and functions of explanations to find essential ingredients of ideal explanation for better user understanding and more accurate mental models [97, 132]. For instance, Lombrozo [133] studied how different types of explanations can help structure conceptual representation. In order to find out how an intelligent system should explain its behavior for non-experts, research on machine learning explanations has studied how users interpret intelligent agents [47, 164] and algorithms [171] to find out what users expect from machine explanations. Related to this, Lim and Dey [124] elicit types of explanations that users might expect in four real-world applications. They specifically study what types of explanations users demand in different scenarios such as system recommendation, critical events, and unexpected system behavior. In measuring user mental model through model failure prediction, Bansal et al. [15] designed a game in which participants receive monetary incentives based on their final performance score. Although experiments were done on a simple three-dimensional task, their results indicate a decrease in users' ability to predict model failure as data and model get more complicated.

A useful way of studying users comprehension of intelligent systems is to directly ask them about the intelligent system's decision-making process. Analyzing users' interviews, think-alouds, and self-explanations provides valuable information about the users' thought processes and mental models [110]. On studying user comprehension, Kulesza et al. [111] studied the impact of explanation soundness and completeness on fidelity of end-users mental model in a music recommendation interface. Their results found that explanation completeness (broadness) had a more significant effect on user understanding of the agent compared to explanation soundness. In another example, Binns et al. [20] studied the relation between machine explanations and users' perception of justice in algorithmic decision-making with different sets of explanation styles. User attention and expectations may also be considered during the interpretable interface design cycles for intelligent systems [195].

Interest in developing and evaluating human-understandable explanations has also led to interpretable models and ad-hoc explainers to measure mental models. For example, Ribeiro et al. [172] evaluated users' understanding of the machine learning algorithm with visual explanations. They showed how explanations mitigate human overestimation of the accuracy of an image classifier and help users choose a better classifier based on the explanations. In a follow-up work, they compared the global explanations of a classifier model with the instance explanations of the same model and found global explanations were more effective solutions for finding the model weaknesses [173]. In another paper, Kim et al. [99] conducted a crowdsourced study to evaluate feature-based explanation understandability for end-users. Addressing understanding of model representations, Lakkaraju et al. [113] presented interpretable decision sets, an interpretable classification model, and measured users' mental models with different metrics such as user accuracy on predicting machine output and length of users' self-explanations.

7.2 M2: Explanation Usefulness and Satisfaction

End-user satisfaction and usefulness of machine explanation are also of importance when evaluating explanations in intelligent systems [19]. Researchers use different subjective and objective measures for understandability, usefulness, and sufficiency of details to assess explanatory value for users [142]. Although there are implicit methods to measure user satisfaction [80], a considerable part of the literature follows qualitative evaluation of satisfaction in explanations, such as questionnaires and interviews. For example, Gedikli et al. [62] evaluated ten different explanation types with user ratings of explanation satisfaction and transparency. Their results showed a strong

Table 4. User satisfaction measures and study methods used in measuring user satisfaction and usefulness of explanations in XAI studies.

Satisfaction Measures	Evaluation Methods
User Satisfaction	Interview and Self-report ([25, 62, 124, 125])
	Likert-scale Questionnaire ([39, 62, 112, 124, 125])
	Expert Case Study ([95, 106, 128, 130, 193])
Explanation Usefulness	Engagement with Explanations ([39])
	Task Duration and Cognitive Load ([62, 112, 125])

relationship between user satisfaction and perceived transparency. Similarly, Lim et al. [125] explore explanation usefulness and efficiency in their interpretable context-aware system by presenting different types of explanations such as "why", "why not" and "what if" explanation types and measuring users response time.

Another line of research studies whether intelligible systems are always appreciated by the users or it has a conditional value. An early work from Lim and Dey [124] studied user understanding and satisfaction of different explanation types in four real-world context-aware applications. Their findings show that, when considering scenarios involved with criticality, users want more information explaining the decision making process and experience higher levels of satisfaction after receiving these explanations. Similarly, Bunt et al. [25] considered whether explanations are always necessary for users in every intelligent system. Their results show that, in some cases, the cost of viewing explanations in diary entries like Amazon and YouTube recommendations could outweigh their benefits. To study the impact of explanation complexity on users' comprehension, Lage et al. [112] studied how explanation length and complexity affect users' response time, accuracy, and subjective satisfaction. They also observed that increasing explanation complexity resulted in lowered subjective user satisfaction. In a recent study, Coppers et al. [39] also show that adding intelligibility does not necessarily improve user experience in a study with expert translators. Their experiment suggests that an intelligible system is preferred by experts when the additional explanations are not part of the translators readily available knowledge. In another work, Curran et al. [41] measured users' understanding and preference of explanations in an image recognition task by ranking and coding user transcripts. They provide three types of instance explanations for participants and show that although all explanations were coming from the same model, participants had different levels of trust in explanations' correctness, according to explanations clarity and understandability.

Table 4 summarizes the study methods used to measure user satisfaction and usefulness of machine learning explanations. Note that the primary goal of XAI system evaluations for domain and AI experts is through direct evaluation of user satisfaction of explanation design during the design cycle. For example, case studies and participatory design are common approaches for directly including expert users as part of the system design and evaluation processes.

7.3 M3: User Trust and Reliance

User trust in an intelligent system is an affective and cognitive factor that influences positive or negative perceptions of a system [82, 136]. Initial user trust and the development of trust over time have been studied and presented with different terms such as *swift* trust [141], *default* trust [139] and *suspicious* trust [21]. Prior knowledge and beliefs are important in shaping the initial state of trust; however, trust and confidence can change in response to exploring and challenging the

Table 5. Evaluation measures and methods used in measuring user trust in XAI studies.

Trust Measures	Evaluation Methods
Subjective Measures	Self-explanation and Interview ([26, 28])
	Likert-scale Questionnaire ([17, 26, 28, 160])
Objective Measures	User Perceived System Competence ([160, 169, 214])
	User Compliance with System ([55])
	User Perceived Understandability ([158, 214])

system with edge cases [81]. Therefore, the user may have different feelings of trust and mistrust during different stages of experience with any given system.

Researchers define and measure trust in different ways. User knowledge, technical competence, familiarity, confidence, beliefs, faith, emotions, and personal attachments are common terms used to analyze and investigate trust [94, 136]. For these outcomes, user trust and reliance can be measured by explicitly asking about user opinions during and after working with a system, which can be done through interviews and questionnaires. For example, Ming et al. [214] studied the importance of model accuracy on user trust. Their findings show that user trust in the system was affected by both the system's stated accuracy and users' perceived accuracy over time. Similarly, Nourani et al. [160] explored how explanation inclusion and level of meaningfulness would affect the user's perception of accuracy. Their controlled experiment results show that whether explanations are human-meaningful can significantly affect perception of system accuracy independent of the actual accuracy observed from system usage. Additionally, trust assessment scales could be specific to the systems application context and XAI design purposes. For instance, multiple scales would assess user opinion on systems reliability, predictability, and safety separately. Related to this, a detailed trust measurement setup is presentation in the paper by Cahour and Forzy [28], which measures user trust with multiple trust scales (constructs of trust), video recording, and self-confrontation interviews to evaluate three modes of system presentation. Also, to better understand factors that influence trust in adaptive agents, Glass et al. [65] studied which types of questions users would like to be able to ask an adaptive assistant. Others have looked at changes to user awareness over time by displaying system confidence and uncertainty of the machine learning outputs in applications with different degrees of criticality [10, 96].

Multiple efforts have studied the impact of XAI on developing justified trust in users in different domains. For instance, Pu and Chen [169] proposed an organizational framework for generating explanations and measured perceived competence and user's intention to return as the measures for user trust. Another example compared user trust with explanations for different goals like transparency and justification explanation [158]. They considered perceived understandability to measure user trust and show that transparent explanations can help reduce the negative effects of trust loss in unexpected situations.

Studying user trust in real-world applications, Berkovsky et al. [17] evaluated trust with various recommendation interfaces and content selection strategies. They measured user reliance on a movie recommender system with six distinct constructs of trust. Also on recommender algorithms, Eiband et al. [55] repeats the Langer et al.'s experiment [114] on the role of "placebic" explanations (i.e., explanations that convey no information) in mindlessness of user behavior. They studied if providing placebic explanations would increase user reliance on the recommender system. Their results suggest that future work on explanations for intelligent systems may consider using placebic

explanations as a baseline for comparison with machine learning generated explanations. Also concerned with expert user's trust, Bussone et al. [26] measured trust by Likert-scale and think-alouds and found that explanations of facts lead to higher user trust and reliance in a clinical decision-support system. Table 5 summarizes a list of subjective and objective evaluation methods to measure user trust in the machine learning systems and their explanations.

Many studies evaluate user trust as a static property. However, it is essential to take user's experience and learning over time into account when working with complex AI systems. Collecting repeated measures over time can help in understanding and analyzing the trend of users' developing trust with the progression of experience. For instance, in their study, Holliday et al. [89] evaluated trust and reliance in multiple stages of working with an explainable text-mining system. They showed that the level of user trust in the system varied over time as the user gained more experience and familiarity with the system.

We note that although our literature review did not find a direct measurement of trust to be commonly prioritized in analysis tools for data and machine learning experts, users' reliance on tools and the tendency to continue using tools are often considered as a part of the evaluation pipeline during case studies. In other words, our summarization is not meant to claim that data experts do not consider trust, but rather we did not find it to be a core outcome explicitly measured in the literature for this user group.

7.4 M4: Human-AI Task Performance

A key goal of XAI is to help end-users to be more successful in their tasks involving machine learning systems [90]. Thus, human-AI task performance is a measure relevant to all three groups of user types. For example, Lim et al. [125] measured users' performance in terms of success rate and task completion time to evaluate the impact of different types of explanations. They use a generic interface that can be applied to various types of sensor-based context-aware systems, such as weather prediction. Further, explanations can assist users in adjusting the intelligent system to their needs. Kulesza et al. [109] study of explanations for a music recommender agent found a positive effect of explanations on users' satisfaction with the agent's output, as well as on users' confidence in the system and their overall experience.

Another use case for machine learning explanations is to help users judge the correctness of system output [70, 105, 194]. Explanations also assist users in debugging interactive machine learning programs for their needs [108, 110]. In a study of end-users interacting with an email classifier system, Kulesza et al. [108] measured classifier performance to show that explanatory debugging benefits user and machine performance. Similarly, Ribeiro et al. [172] found users could detect and remove wrong explanations in text classification, resulting in training better classifiers with higher performance and explanations quality. To support these goals, Myers et al. [156] designed a framework that users can ask *why* and *why not* questions and expect explanations from the intelligent interfaces. Table 6 summarizes a list of evaluation methods to measure task performance in human-AI collaboration and model tuning scenarios.

Visual analytics tools also help domain experts to better perform their tasks by providing model interpretations. Visualizing model structure, details, and uncertainty in machine outputs can allow domain experts to diagnose models and adjust hyper-parameters to their specific data for better analysis. Visual analytics research has explored the need for model interpretation in text [92, 128, 210] and multimedia [24, 33] analysis tasks. This body of work demonstrates the importance of integrating user feedback to improve model results. An example of a visual analytics tool for text analysis is TopicPanaroma [131], which models a textual corpus as a topic graph and incorporates machine learning and feature selection to allow users to modify the graph interactively. In their evaluation procedure, they ran case studies with two domain experts: a public

Table 6. Evaluation measures and methods used in measuring human-machine task performance in XAI studies.

Performance Measures	Evaluation Methods
User Performance	Task Performance ([70, 95, 110, 125])
	Task Throughput([110, 113, 125])
	Model Failure Prediction ([70, 105, 194])
Model Performance	Model Accuracy ([108, 130, 165, 172, 194])
	Model Tuning and Selection ([131])

relations manager used the tool to find a set of tech-related patterns in news media, and a professor analyzed the impact of news media on the public during a health crisis. In analysis of streaming data, automated approaches are error-prone and require expert users to review model details and uncertainty for better decision making [18, 179]. For example, Goodall et al. [66] presented Situ, a visual analytics system for discovering suspicious behavior in cyber network data. The goal was to make anomaly detection results understandable for analysts, so they performed multiple case studies with cybersecurity experts to evaluate how the system could help users to improve their task performance. Ahn and Lin [4] present a framework and visual analytic design to aid fair data-driven decision making. They proposed FairSight, a visual analytic system to achieve different notions of fairness in ranking decisions through visualizing, measuring, diagnosing, and mitigating biases.

Other than domain experts using visual analytics tools, machine learning experts also use visual analytics to find shortcomings in the model architecture or training flaws in DNNs to improve the classification and prediction performance [130, 165]. For instance, Kahng et al. [95] designed a system to visualize instance-level and subset-level of neuron activation in a long-term investigation and development with machine learning engineers. In their case studies, they interviewed three machine learning engineers and data scientists who used the tool and reported the key observations. Similarly, Hohman et al. [86] present an interactive system that scalably summarizes and visualizes what features a DNN model has learned and how those features interact in instance predictions. Their visual analytic system presents activation aggregation to discover important neurons and neuron-influence aggregation to identify interactions between important neurons. In the case of recurrent neural networks (RNN), LSTMVis [193] and RNNVis [143] are tools to interpret RNN models for natural language processing tasks. In another recent paper, Wang et al. [206] presented DNN Genealogy, an interactive visualization tool that offers a visual summary of DNN representations.

Another critical role of visual analytics for machine learning experts is to visualize model training processes [224]. An example of a visual analytics tool for diagnosing the training process of a deep generative model is DGMTracker [129], which helps experts understand the training process by visually representing training dynamics. An evaluation of DGMTracker was conducted in two case studies with experts to validate efficiency of the tool in supporting understanding of the training process and diagnosing a failed training process.

7.5 *M5: Computational Measures*

Computational measures are common in the field of machine learning to evaluate interpretability techniques' correctness and completeness in terms of explaining what the model has learned. Herman [78] notes that reliance on human evaluation of explanations may lead to persuasive explanations rather than transparent systems due to user preference for simplified explanations.

Table 7. Evaluation measures and methods used for evaluating fidelity of interpretability techniques and reliability of trained models. This set of evaluation methods is used by machine learning and data experts to eighter evaluate the correctness of interpretability methods or evaluate the training quality trained models beyond standard performance metrics.

Computational Measures	**Evaluation Methods**
Explainer Fidelity	Simulated Experiments ([172, 173])
	Sanity Check ([101, 162, 177, 215, 217, 226]
	Comparative Evaluation ([178, 183])
Model Trustworthiness	Debugging Model and Training ([219])
	Human-Grounded Evaluation ([43, 134, 149, 187])

Therefore, this problem leads to the argument that explanations' fidelity to the black-box model should be evaluated by computational methods instead of human subject studies. Fidelity of an ad-hoc explainer refers to the correctness of the ad-hoc technique in generating the true explanations (e.g., correctness of a saliency map) for model predictions. This leads to a series of computational methods to evaluate correctness of generated explanations, consistency of explanation results, and fidelity of ad-hoc interpretability techniques to the original black-box model [175].

In many cases, machine learning researchers often consider consistency in explanation results, computational interpretability, and qualitative self-interpretation of results as evidence for explanation correctness [162, 215, 217, 226]. For example, Zeiler and Fergus [219] discuss fidelity of the visualization for CNN network by its validity in finding model weaknesses resulted in improved prediction results. In other cases, comparing a new explanation technique with existing state-of-the-art explanation techniques is is used to verify explanation quality [37, 134, 189]. For instance, Ross et al. [178] designed a set of empirical evaluations and compared their explanations' consistency and computational cost with the LIME technique [172]. In a comprehensive setup, Samek et al. [183] proposed a framework for evaluating saliency explanations for image data that quantify the feature importance with respect to the classifier prediction. They compared three different saliency explanation techniques for image data (sensitivity-based [190], deconvolution [219], and layer-wise relevance propagation [13]) and investigated the correlation between saliency map quality and network performance on different image datasets under input perturbation. On the contrary, Kindermans et al. [101] show interpretability techniques have inconsistencies on simple image transformations, hence their saliency maps can be misleading. They define an input invariance property for reliability of explanations from saliency methods. To extend a similar idea, Adebayo et al. [3] propose three tests to measure adequacy of interpretability techniques for tasks that are sensitive to either data or the model itself.

Other evaluation methods include assessing explanation's fidelity in comparison to inherently interpretable models (e.g., linear regression and decision trees). For example, Ribeiro et al. [172] compared explanations generated by the LIME ad-hoc explainer to explanations from an interpretable model. They created gold standard explanations directly from the interpretable models (sparse logistic regression and decision trees) and used these for comparisons in their study. A downside of this approach is that the evaluation is limited to generating a gold standard by an interpretable model. User simulated evaluation is another method to perform computational evaluation of model explanations. Ribeiro et al. [172] simulated user trust in explanations and models by defining "untrustworthy" explanations and models. They tested a hypothesis on how real users

would prefer more reliable explanations and choose better models. The authors later repeated similar user simulated evaluations in the *Anchors* explanation approach [173] to report simulated users' precision and coverage in finding the better classifier by only looking at explanations.

A different approach in quantifying explanations quality with human intuition has been taken by Schmidt and Biessmann [187] by defining an explanation quality metric based on user task completion time and agreement of predictions. Another example is the work of Lundberg and Lee [134], who compared the SHAP ad-hoc explainer model with LIME and DeepLIFT [189] with the assumption that good model explanations should be consistent with the explanations from humans who understand the model. Lertvittayakumjorn and Toni [118] also present three user tasks to evaluate local explanation techniques for text classification through revealing model behavior to human users, justifying the predictions, and helping humans investigate uncertain predictions. A similar idea has been implemented in [149] by feature-wise comparison of a ground-truth and model explanation. They provide a user annotated benchmark to evaluate machine learning instance explanations. Later, Poerner et al. [166] use this benchmark as human annotated ground truth in comparison to small-context (word level) and large-context (sentence level) explanation evaluation. User annotated benchmarks can be valuable when considering human meaningfulness of explanations, though the discussion by Das et al. [43] implies that machine learning models (visual question answering attention models in their case) do not seem to look at the same regions as humans. They introduce a human-attention dataset [42] (collection of mouse-tracking data) and evaluate attention maps generated by state-of-the-art models against human.

Interpretability techniques also enable quantitative measures for evaluating model trustworthiness (e.g., model fairness, reliability, and safety) through its explanations. Trustworthiness of a model represents a set of domain specific goals such as fairness (by fair feature learning), reliability, and safety (by robust feature learning). For example, Zhang et al. [220] present a case of using machine learning explanations to find representation learning flaws caused by potential biases in the training dataset. Their technique mines the relationships between pairs of attributes according to their inference patterns. Further, Kim et al. [99] presented quantitative testing of machine learning models by their explanations. In their concept activation vector technique, the model can be tested for specific concepts (e.g., image patterns) and a vector score shows if the model is biased toward that concept. They later extended their concept-based global explanation of model representation learning for systematic discovery of concepts that are human meaningful and important for the model prediction [63]. They used human subject experiments to evaluate learned concepts. Table 7 summarizes a list of evaluation methods to measure fidelity of interpretability techniques and model trustworthiness with computational techniques.

8 XAI DESIGN AND EVALUATION FRAMEWORK

The variety of different XAI design goals (Section 6) and evaluation methods (Section 7) from our review suggests the need for diverse sets of techniques to build end-to-end XAI systems. However, it is generally insufficient to take design practices and evaluation methods separately. A holistic and more actionable vantage will require consideration of dependencies between design goals and evaluation methods and will inform when to choose between them during the design cycles. Previously, various models and guidelines for the design and evaluation of AI-infused interactive user interfaces [7, 54] and visual analytics systems [155] have been proposed to help designers through the design process. However, challenges in generating useful machine learning explanations and presenting them through an appropriate interface that aligns with target outcomes call for a multidisciplinary workflow framework.

Thus, based on our analysis of prior work, we propose a design and evaluation framework for XAI systems. The impetus for this framework is the desire to organize and relate the diverse set of

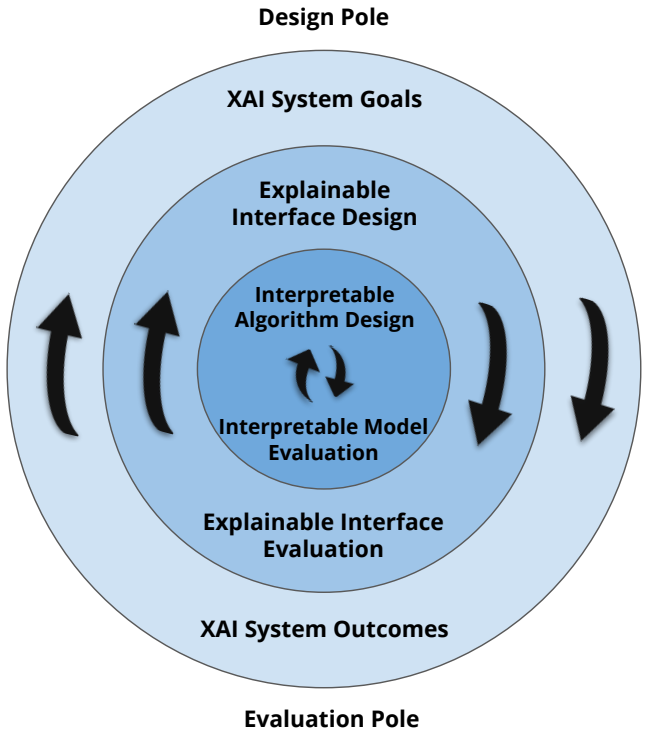

Fig. 4. XAI Design and Evaluation Framework: our nested model for design and evaluation of explainable machine learning systems. The *outer layer* demonstrates system-level design goals which are paired with evaluation of high-level XAI outcomes. The *middle layer* shows explainable user interface and visualization design step paired with appropriate user understandability and satisfaction evaluation measures. The *innermost layer* presents design and evaluation of trustworthy interpretable machine learning algorithms.

existing design goals and evaluation methods in a unified model. The framework is intended to give guidance on what evaluation measures are appropriate to use at which design stage of the XAI system. Figure 4 summarizes the framework as a nested model for end-to-end XAI system design and evaluation. The formulation of the model as layers relates to the core design goals and evaluation interests from the different research communities (as identified from the literature review) to help promote interdisciplinary progress in XAI research. The model is structured to support system design steps by starting from the outer layer (*XAI System Goals*), then addressing end-user needs in the middle layer (*Explainable Interface*), and finally focusing on underlying interpretable algorithms in the innermost layer (*Interpretable Algorithms*). The nested model is organized with a *Design Pole* focusing on design goals and choices, and an *Evaluation Pole* presenting appropriate evaluation methods and measures for each layer. Our framework suggests iterative cycles of design and evaluation to cover both algorithmic and human-related aspects of XAI systems. In this section, we elaborate on details of the nested framework and provide guidelines on using it for multidisciplinary XAI system design.

Case Study Example: To showcase a practical example of using the framework, we also include a case study of a collaborative design and development effort for an XAI system. In the scenario of the case study, a multidisciplinary team of researchers designed a XAI system for fake news

detection for non-expert (not AI experts or news analysts) daily newsreaders. The design team planned to add a *XAI Assistant* feature to a news reading and sharing website to perform fake news detection. The system design consisted of a news reading interface equipped with the XAI news assistant (news assistant) to help the user identify fake news while reviewing news stories and articles. The presented case is the result of an ongoing research done over a one-year period by a team of eight university researchers with HCI, Visualization, and AI backgrounds. During the following subsections, each framework guideline is followed by an example of its application in our case study.

8.1 XAI System Goals Layer

As designers in a multidisciplinary team have different roles and priorities in building a XAI system, we suggest beginning the system design cycle from the *XAI Goal* layer (the outer layer of Figure 4) to characterize design goal and system expectations. Specifically, this step involves identifying the *purpose for explanation* and choosing *what to explain* for the targeted end-user and dedicated application. The iterative refinements between XAI goal (top pole) and system evaluation (bottom pole) present how the paired evaluation measures help to improve system design. We organize the following guidelines for the XAI goal layer.

At the beginning of the system design process, the team will need to specify explainability requirements for each framework layer based on the evaluation metrics. The explainability requirements are intended to satisfy overall system goals defined by user (or customer) needs, and sometimes regulations, laws, and safety standards. Later, the evaluation step in each design cycle will have the team revisit the initial XAI system requirements. The sufficiency of the evaluation results in comparison to the initial design requirements serves as a key indicator of whether to stop or continue design iteration. However, since many subjective measures are used in the process, it is important to choose an appropriate evaluation baseline (see Section 9.4) to track progress during design cycles.

Guideline 1: Determine XAI System Goals: Identifying and establishing clear goals and expectations from XAI system is the first step in the design process. XAI Design goals could be driven by many motivations like improving user experience on an existing system, advancing scientific findings [107, 120], or adhering to new regulations [198]. In Section 6 we reviewed eight main goals (G1-G8) for XAI systems. Also, ordering the priority of goals in cases with multiple design goals can be beneficial in the next steps of the process (see Guideline 2). Given the fact that different XAI user types and applications are interested in various design goals, it is important to establish these goals early in the design process to identify and align with appropriate design principles. A pitfall in this stage is to pick XAI goals without considering the end-user group, algorithmic limitations, and user preferences in the context of the application. Overshooting XAI goals could hurt evaluation results moving forward in the design process.

Application in Case Study: In the first step of our case study with a news curation application, the team started with identifying the main goals and expectations for the XAI news assistant. The design focused on novice end-users without any particular expertise. The XAI design goal was to improve user reliance and mental model of news predictions through explainable design. The team hypothesized that end-users would trust and rely on the fake news detection assistant, given that the new XAI is capable of providing explanations for each news story. Also, the team hoped that users would be able to use the explanations to learn model weaknesses and strengths to provide feedback to the developer team.

Guideline 2: Decide What to Explain:

The second step in the XAI system design is to identify "what to explain" to the user in order to achieve the initial XAI goals (see Guideline 1) of the system. We reviewed multiple machine learning interpretability techniques and explanation types in Sections 4.1, 4.2, and 4.3 which can provide different types of information to the user. Although theory-driven design frameworks discuss explanation mechanisms driven by human reasoning semantics [126], user-centered methods to identify useful explanations such as online surveys, interviews, and user observations (e.g., [26, 138]) to understand *when* and *what* needs to be explained for the users to understand better and trust intelligent systems. Preliminary experiments are valuable in the early steps of the design cycle to identify and narrow down explanation options for the user in order to satisfy design goals. A typical approach for evaluating the effectiveness and usefulness of explanation choice in user-centric experiments is to compare the user's mental model of the system with and without explanation components. On this subject, Lim and Dey [124] conducted experiments to discover what type of information users are interested in different real-world context-aware application scenarios. Stumpf et al. [195] also performed end-user interviews to identify user perceptions and expectations from an interpretable interface as well as finding main decision points where users may need explanations. In another work, Haynes et al. [74] provide a review and studies incorporating different explanations (operational, ontological, mechanistic, and design rationale explanations) in intelligent systems. Similarly, visual analytics design involves expert interviews and focus groups in the design path to identify design goals [155].

The design process in this step involves algorithmic implementation constraints like "what can be explained" to the user. For example, global explanations from a DNN may not feasible and comprehensible due to the large number of variables in the graph. Additionally, research shows instance explanations from a DNN lack completeness and may fail to present salient features in cases [3]. Such constraints and decision points could be solved through focused groups, brainstorming, and interviews between model designers and interface designers in the team. Therefore, a design pitfall for explanation choices is not to take limitations of interpretability techniques into account.

> *Application in Case Study:* In our scenario, efficient news curation required fake news detection with the help of our XAI assistant. In the analysis of what the system should explain, the design team decided to identify candidate useful and impactful explanation options. We started with reviewing machine learning research on false information (e.g., rumor, hoax, fake news, clickbait) detection as well as HCI research on news feeds and news search systems to identify key attributes for news veracity checking [148]. Given the non-expert target end-users, explanatory information needed to limit technical details. Next, the user interface designers and machine learning designers in the team discussed candidate explanation choices and algorithmic constraints in interpretability techniques. That is, some options for what to explain may not be entirely possible given the interpretability of existing models, and the team needed to consider whether alternative learning techniques could provide better explanations or if the design team would need to figure out meaningful ways to explain the information that was available from the model.

Guideline 3: Evaluate System Outcomes: Evaluation of XAI system outcomes is the final step in the evaluation process. Figure 4 shows how the final system outcome evaluation is paired with the initial design goals in the outer layer of our framework. The main goal of this stage is to quantitatively and qualitatively assess the effectiveness of the XAI system for the initially established system-level XAI goals. Clearly, evaluation of final system outcomes could be influenced

by the design of the explainable user interface (intermediate layer) and the design of interpretable algorithms (innermost layer). For example, evaluating a newborn interpretable machine learning algorithm's output using human subjects through a weak in-lab or crowdsourced user study may not be meaningful or productive for XAI system outcomes if core computational changes are still in progress and could ultimately change the entire model interpretability and explanation format later. Also, changes in the targeted user could affect evaluation results at this stage. For example, a system designed for novices may not satisfy the needs of an expert user and hence would not improve performance as expected. Evaluation measures in this layer depend on the design goals, application domain, and targeted users. Example evaluation measures for final system outcomes include user trust [169] and reliance on the system [17], human-machine task performance [15], user awareness [96], and user understanding of their personal data [170]. An effective process for evaluation of high-level XAI outcomes is to break down the evaluation goal into multiple well-defined measures and metrics. This way, the team can perform evaluation studies on different steps using valid methods in controlled setup. For example, in the evaluation of XAI systems for trustworthiness, several factors of human trust could be measured during and after a period of user experience with the XAI system. In addition, computational measures (Section 7.5) are used to examine the fidelity of interpretability methods and trustworthiness of the model with objective metrics. A possible pitfall in evaluation of the XAI system outcomes is performing the evaluation without considering the model trustworthiness and explanations' correctness from the interpretable model layer (see Guideline 7) and explanation understandability and usefulness from the user interface layer (see Guideline 5).

Application in Case Study: In our case study with news review and curation, we needed to evaluate our XAI news assistant with non-expert users who would gather news stories while flagging fake news articles. In the evaluation step, the team ran multiple large-scale human-subject studies with novice participants recruited through Amazon Mechanical Turk to work with our news reading system. Note that both the explainable interface and interpretable algorithm passed multiple design and testing iterations before this evaluation step. Major decisions for this evaluation was how to structure the duration and complexity of the user task while appropriately testing the system's full range of functionality. The task was designed with questions built in to help collect subjective data in addition to the objective user performance data. Multiple evaluation measures were chosen for system outcomes, including: subjective user trust in the news assistant, user agreement rate with the news assistant, veracity of user-shared news stories, and user accuracy in guessing the news assistant output. Both qualitative and quantitative analysis of user feedback and interaction data were valuable to the evaluation of system outcomes. The results and analysis from these evaluations helped the team to understand the effectiveness of the XAI elements (in both the algorithm and the interface) for the initial system goals.

8.2 User Interface Design Layer

The middle layer of our framework is concerned with designing and evaluating an explainable interface or visualization for the user to interact with XAI system. Interface design for explanations consists of presenting model explanations from interpretable algorithms to end-users in terms of their *explanation format* and *interaction design*. The importance of this layer is to satisfy design requirements and needs to be determined in the XAI system design layer (see Guideline 2). An

elegant translation of machine-generated explanations (e.g., verbal, numeric, or visual explanation) needs carefully designed human-understandable and satisfying explanations in the user interface. In Section 4.4, we reviewed multiple types of explanation formats for integrating XAI elements into the user interface. The iterative movement between *Design pole* and *Evaluation pole* in this layer presents design refinement in pursuit a desired goal state.

Guideline 4: Decide How to Explain:
Identifying candidate explanation formats for the targeted system and user group is the first step to deliver machine learning explanations to end-users. The design process can account for different levels of complexity, length, presentation state (e.g., permanent or on-demand), and interactivity options depending on the application and user type. The explanations format in the interface is particularly important to improve user understanding of underlying algorithms. Studies show that while detailed and complex interactive representations may aim to communicate the explanations to the expert users, AI-novice users of XAI system prefer more simplified explanation and representation interfaces [112]. User satisfaction of interface design is also another critical factor in user engagement of the interface components [154]. Additionally, interaction design for explainable interfaces can allow a user to communicate with the system to adjust explanations and could better support user inspection of the system [110]. Research of intelligent interface design presents multiple design methods such as wireframing and low-fidelity prototyping (e.g., [26, 138]) that could also be adapted to the explainable interface design. Also, existing design guidelines and best-practice knowledge for AI-infused interfaces (e.g., [7]) and visualizations (e.g., [140]) could be used in this stage to leverage similar systems for explainable interface design. Aside from model explanations, providing prediction uncertainty also has been identified as an important factor for both general end-users and data expert users [181]. For example, Kay et al. [96] presented the full design cycle for an uncertainty visualization interface in a bus arrival time application. Their design process included surveying to identify usage requirements, developing alternative layouts, running user testing, and final evaluation of user understanding of machine learning output.

> *Application in Case Study:* To determine *how* to explain news classification results to non-expert end users, the user interface design team started the process by reviewing the initial system goals and explanation types. The team then continued with multiple interface sketches that matched the intended application and user tasks. During the initial design steps, the team tried to keep a balance between interface complexity and explanation usefulness by choosing among available explanation types from our interpretable machine learning algorithms. Next, mock-ups from the top three designs were implemented for testing with a small number of participants. Each mock-up had a different arrangement of data, user task flow, and explanation format for the news assistant interface. Our human-subject experiments in this stage were based on user observations and post-usage interviews to collect qualitative feedback regarding participant understanding and subjective satisfaction of explanation components and interface arrangements. Interviews resulted in the selection of the most comprehensible and conclusive design among the available options to continue with (see Guidelines 5).

Guideline 5: Evaluate Explanation Usefulness: This mid-layer evaluation step can be used along with various measures to help assess user understanding of the XAI underlying intelligent algorithms. A series of user-centered evaluations of explainable interface with multiple goals and granularity levels could be performed to measure:

(1) User understanding of explanation.
(2) User satisfaction of explanation.
(3) User mental model of the intelligent system.

Evaluations in the middle layer are particularly important due to the impact on XAI system outcomes (outer layer) and being affected by interpretable model outputs (inner-most layer). Specifically, evaluation measures in this stage can inform how well users understand the interpretable system, however, the design validity at this step also may be reflected by higher-level XAI outcomes (i.e., outer-layer evaluation) such as user trust and task performance. Note that user understanding of an XAI system could be limited to parts of the system rather than the entire system; similarly, understanding may be limited to a subspace of scenarios rather than all possible scenarios.

The three evaluation measures introduced for this step could be used on multiple iterative cycles to improve overall explainable interface design. For example, Saket et al. [182] studies users understanding of visualization encoding and effectiveness of interactive graphical encoding for end-user. On the other hand, user satisfaction of explanation type and format depends on factors such as targeted application criticality and user-preferred cognitive load [48]. Evaluating user mental model is also an effective way to measure usefulness of explainable interfaces. Tables 3 and 4 present a list of measures for evaluating explainable interfaces in this step. The choice of baseline is another important factor in evaluating explainable interfaces. Typically, a combination of qualitative and quantitative analysis are used to measure effects of explanation components (in comparison to non-explainable system) or to compare multiple explanations types. However, the choice of placebic explanations has been proposed as the evaluation baseline for more accurate measurement of explanation content [55]. In the case of expert review, evaluation of a domain expert's mental model commonly involves comparison with the AI expert's mental model and description of "how the model works". Section 9.4 reviews common choices of ground-truth baselines in XAI evaluation studies. With all approaches, updates in explanation components of the interface require assessment of their impact on user experience and understandability. However, the metrics and depth of evaluation vary during the evaluation cycles as the team narrows down specific needs. Finally, a possible evaluation pitfall for explainable interfaces is going after broad measures of XAI outcomes (See Guideline 3) rather than focusing on a narrower scope of explanation components and interactions.

Application in Case Study: In our case study, interface designers started evaluation of candidate explanation components by a series of small studies with a repeated-measures design so that the same study participant could experience different explanation designs in one session. Next, we analyzed quantitative and qualitative data collected from the end-users to choose candidate designs and routes to further improve the interface for explainable components. Discussions with the machine learning team also helped to find sources of limitations in the interpretability technique that could possibly affect user satisfaction. After the initial cycles of revision, we collected a round of external and internal expert reviews to update the study methodology and data collection details according to project progress.

8.3 Interpretable Algorithm Design Layer

The innermost layer of our framework involves designing interpretable algorithms that are able to generate explanations for the users. The last design step in our XAI system framework is the

choice of interpretability technique (design pole) to generate the outlined explanation types. However, evaluating the generated explanation (evaluation pole) is the first evaluation step before human-subject evaluations in the explainable interface. Ideally, the interpretability techniques should generate explanations in accordance with the requirements in the explainable interface design step (see Guideline 4); however, the choice of interpretability technique depends on domain and carries implementation limitations. For example, while shallow models are desired for their high interpretability, these models typically do not perform well in cases of complex and high dimensional data like image and text. On the other hand, highly accurate predictions in black-box models (e.g., deep neural networks and random forest models) require post-processing and ad-hoc algorithms to generate explanations. The ad-hoc approach also has limitations on both choice of explanation type and need for completeness [3] and fidelity [172] validation compared to the original model. This shows not only machine learning designers should consider the trade-off between model interpretability and performance but also should consider the fidelity of the ad-hoc explainer to black-box model. We suggest two following design and evaluation guidelines for this layer:

Guideline 6: Design Interpretability Technique: Designing interpretable decision-making algorithms starts with the choice of machine learning model. Shallow machine learning models (e.g., linear models and decision trees) have intrinsic interpretability due to low number of variables and model simplicity. For more complex models (e.g., random forest and DNN), ad-hoc explainer technique (see Section 4.2) are needed to generate explanations. However, the choice of machine learning model (i.e., shallow vs. deep) is bounded by model's performance on data domain. Secondly, ad-hoc explainer techniques have certain limitations in their explanation type. The importance of choosing the right combination of model and explainer is in their impact on providing useful (See Guideline 4) and trustworthy explanations for end-users.

Machine learning research has proposed various ad-hoc explainers to generate "Why" explanations (e.g., feature attribution [99, 134]), "How" explanations (e.g., rules list [119, 205]), "What else" explanation (e.g., similar training instance [98, 135]), and "What if" (e.g., sensitivity analysis [219]) explanation types. However, despite substantial research in interpretable machine learning techniques, a core issue in model explanations is the difference between machine learning model's decision-making logic and human sense-making as the receiver [100, 223].

Application in Case Study: In our fake news detection case study, the explainable interface design team had previously discussed candidate explanation choices with the machine learning design team (see Guidelines 2 and 4). Therefore, a final review of model-generated explanations and an assessment of implementation limitations were performed in this step. For example, removing noise-like features from saliency maps, normalizing attributions scores, and resolving contradicting explanations between an ensemble of models were primary implementation bottlenecks that were resolved in this step. Specifically, as a decision point for trade-offs between clarity and faithfulness of explanations, the team decided on using heuristic filters to eliminate features with a very low attribution score for the sake of presentation simplicity.

Guideline 7: Evaluate Model Trustworthiness: Evaluating the interpretable machine learning is the first evaluation step in our framework due to its impact on outer layer evaluation measure. The high significance of this evaluation step stems from the possibility that any unreliability of interpretability at this inner layer will propagate to all other outer layers. Such unintended error

propagation may lead to problematic outer-layer design decisions as well as misleading evaluation results. We discuss two main evaluation goals for the innermost layer:

(1) Evaluating model trustworthiness.
(2) Evaluating ad-hoc explainer fidelity.

The first evaluation goal aims to utilize interpretability techniques as a debugging tool to analyze the model's trustworthiness on learning concepts beyond general performance measures [99]. Examples of model trustworthiness validation include evaluating model reliability in financial risk assessment [60], model fairness in social influencing applications [220], and model safety for its intended functionality [22]. Researchers have also proposed various regularization techniques for enhancing trustworthy feature learning in machine learning models [76, 178]. Next, the second evaluation goal targets fidelity of ad-hoc explainer techniques to the black-box model. Research shows that different ad-hoc interpretability techniques have inconsistencies and can be misleading [3]. Evaluating explanation trustworthiness can verify explainer fidelity in terms of how well it represents the black-box model (see Section 7.5).

Application in Case Study: In our case study, we paid careful attention to qualitative reviewing of the model explanations after each design iteration. Our initial qualitative review of model explanations led to dataset cleaning through a heuristic search aimed at the removal of mislabeled examples and unrelated news articles. An improvement to model performance was achieved after dataset cleaning. Then, after the first round of human-subject evaluation of the explainable interface (see Guideline 5), the team identified negative effects of keyword explanations with low attention scores from end-users. The team decided on using a lower threshold for visualizing attention maps to reduce clutter and "noisy explanations" for end-users. Finally, after one round of XAI outcome evaluation (see Guideline 3), analysis of users' mental models revealed that a dataset imbalance between the "fake news" and "true news" was causing a bias for the model in that the model was usually more confident in predicting fake news over true news.

9 DISCUSSION

In our review, we discussed multiple XAI design goals and evaluation measures appropriate for various targeted user types. Table 2 presents the categorization of selected existing design and evaluation methods that organizes literature along three perspectives: *design goals*, *evaluation methods*, and the *targeted users* of the XAI system. Our categorization revealed the necessity of an interdisciplinary effort for designing and evaluating XAI systems. To address these issues, we proposed a design and evaluation framework that connects design goals and evaluation methods for end-to-end XAI systems design, as presented through a model (Figure 4) and guidelines. In this section, we discuss further considerations for XAI designers to benefit from the body of knowledge of XAI system design and evaluation. The following recommendations support and promote different layers of the proposed design and evaluation framework as well.

9.1 Pairing Design Goals with Evaluation Methods

It is essential to use appropriate measures for evaluating the effectiveness of design elements. A common pitfall in choosing evaluation measures in XAI systems is that the same evaluation measure is sometimes used for multiple design goals. A simple solution to address this issue is to distinguish between measurements by using multiple scales to capture different attributes in each

evaluation target. For example, the concept of *user trust* consists of multiple constructs [28] that could be measured with separate scales in questionnaires and interviews (see Section 7.3). User satisfaction measurements could also be designed for various attributes such as understandability of explanations, usefulness of explanations, and sufficiency of details [85] to target specific explanation qualities (see Section 7.2).

An efficient way to pair design goals with appropriate evaluation measures is to balance different design methods and evaluation types in iterative cycles of design. Managing the trade-offs between qualitative and quantitative methods in the design process can allow designers to take advantage of different approaches, as needed. For example, while focus groups and interviews provide more detailed and in-depth feedback on the users' mental model [109], remote measurements are highly valuable due to the scalability of the collected data even though they provide less detail for drawing conclusions [112]. Thus, one successful approach could be to start with multiple small-scale prototyping and formative studies collecting qualitative measures at the earlier stages of the design (e.g., for XAI system goals layer in the framework) and continue with larger-scale studies and quantitative measures in the later stages (e.g., for interpretable model and interface evaluations in the framework).

9.2 Overlap Among Design Goals

In our categorization of XAI systems, we chose two main dimensions to organize XAI systems by their *Design Goals* and *Evaluation Measures* in Section 5. The XAI design goals (G1–G8) were based on the goals extracted from the surveyed papers, and since the XAI design goals are primarily derived from their targeted user groups, we note that overlaps among goals do exist across disciplines. For instance, there is overlap of the goals of *G1: Algorithmic Transparency* for novice users in HCI research, *G5: Model Visualization* for data experts in visual analytics, and *G7: Interpretability Techniques* for AI experts in machine learning research. While overlapping, these similar goals are studied with different objectives across the three research disciplines leading to diverse sets of design requirements and implementation paths. For example, designing XAI systems for AI novices requires processes and steps to build human-centered explainable interfaces to communicate model explanations to the end-users, whereas designing new interpretability techniques for AI experts has a different set of computational requirements. Another example of overlap in XAI goals is between the goal for *G6: Model Visualization and Inspection* for data experts and *G8: Model Debugging* for AI experts, in which different sets of tools and requirements are used to address different research objectives.

To address the overlap between XAI goal among research disciplines, we used the XAI *User Groups* as an auxiliary dimension to organize XAI goals in this cross-disciplinary topic (Section 6) and emphasize the diversity of diverse research objectives. The three user groups were chosen to organize research objectives and efforts into HCI (for AI novices), visual analytics (for data experts), and machine learning (for AI experts) research fields. Additionally, as described in the framework, the three user groups prioritize design objectives in the design process for the XAI system rather than absolute separation of design goals. For example, the objectives and priorities in XAI system design for algorithmic bias mitigation for domain experts in a law firm are certainly different from those of model training and tuning tools for AI experts. However, by following the multidisciplinary design framework, a design team can translate XAI system goals into design objectives for explainable interface and machine learning techniques to improve the design process in different layers. Therefore, in the above example, the design team can focus on diverse interface design and interpretability technique objectives to achieve the primary XAI goal of bias mitigation for the domain experts. Note that the specifics of any particular system will determine the priorities of different objectives.

9.3 System Evaluation Over Time

An important aspect in evaluating complex AI and XAI systems is to take the user learning into account. Learnability is even more critical when measuring mental models and user trust in the system. A user learns and gets more familiar with the system over time with continued interaction with the system. This brings the importance of repeated temporal data capture (in contrast to static measurements) for XAI evaluations. Holliday et al. [89] present an example of multiple trust assessments during the user study. They measured user trust at regular intervals during the study to capture changes in user trust as the user interacts more with the system. Their results indicates an XAI system outperformed a non-XAI counterpart in maintaining user trust over time. Time-based measurements, also referred to as *dynamic measurements*, allows designers to monitor explanation usability and effectiveness in various contexts and situations [50, 185]. For instance, Zhang et al. [222] explore the effect of intelligent system explanations in user trust calibration. In their experiments, they observe significant effect on calibration of trust when model prediction confidence score was shown to participants. In another example, a study by Nourani et al.[159] controlled whether users' early experiences with an explainable activity recognition system had better or worse model outputs, and the first impressions significantly affected both task performance and user confidence in understanding how the system works. In a study with a news review task, Mohseni et al. [150] identified different user profiles for changes in trust over time (trust dynamics) while working with the assistance of an explainable fake news detector. Their analysis of results revealed a significant effect of machine learning explanations on user trust dynamics.

Long-term evaluation of XAI systems can also allow designers to estimate valuable user experience factors such as over-trust and under-trust on the system. User-perceived system accuracy [110] and transparency [171] are examples of long-term measures for explanation usability that depend on building user trust in the system's interpretability. As more information is provided by explanations over time, reasoning and mental strategies may change as users create new hypotheses about system functionality. Therefore, it is essential to also consider users' mental models and trust in extended studies to evaluate all aspects of the XAI system.

Another use case of long-term measurements is to evaluate the effects of intelligent system's non-uniform behaviors in real-world scenarios. This means, although in a controlled study setup, a balanced set of input examples will present the system to the user, in real-world scenarios, users may face alterations in system performance in long-term interaction with the system. Long-term measurements will identify user's unjust trust in the system due to a limited or biased set of interactions with the system. For example, in the context of autonomous vehicles, Kraus et al. [104] presented a model of trust calibration and presented studies on trust dynamics in the early phases of user interaction with the system. Their results indicate the effects of error-free automation in steady increase of user trust as well as the effects of user a priori information in eliminating the decrease of trust in case of system malfunction.

9.4 Evaluation Ground Truth

Research on XAI systems study various goals with different measures across multiple domains. The breadth of XAI research makes it challenging to interpret and transfer findings from one task and domain to another. Knowing key factors for interpreting implications of evaluation results is essential to aggregate findings across domains and disciplines. An important factor in understanding XAI evaluation results and comparing results among multiple studies is the choice of ground truth. In the following, we review common choices of ground truth for both human-subject and computational evaluation methods.

Human-subject experiments often take the form of controlled studies to examine the effects of machine learning explanations on a control group in comparison to a baseline group. In these setups, the choice of the baseline could affect results implications and significance. Our review of papers in the space of XAI evaluation shows the majority of study designs use a *no explanation* condition as the baseline condition to measure the effectiveness of model explanations in an explanation group. Examples for the baseline include approaches that remove model explanations related components and features form the interface in the baseline condition [103, 160]. In other work, Poursabzi et al. [168] also included a *no AI* baseline to measure participants' performance without the help of model predictions. Another way is to compare the effects of explanation type or complexity between study conditions without the *no explanation* baseline. For instance, Lage et al. [112] present a study to evaluate the effects of explanation complexity on participants' comprehension and performance. They used linear and logistic regression to estimate the effects of explanation complexity on participants' normalized response time, response accuracy, and subjective task difficulty rating.

Though the above mentioned studies are controlled experiments, there may still be unaccounted human behavioral implications due to differences in the complex process of explaining worthy of consideration. Langer et al. [114] present an experiment on "placebic" explanations that shows people's mindless behavior when facing explanations for actions. In a simple setup, their study showed that when asking a request, inclusion of explanations and justifications increased user's willingness to comply even if the explanations convey no meaningful information. Recently, Eiband et al. [55] proposed using *placebic explanations* instead of a *no explanation* condition as the baseline for XAI human subject studies. Therefore, using non-informative or even randomly generated explanations as the baseline condition could potentially counteract a participant's positive tendency toward explanations and improve study results.

Considering other approaches, a commonly accepted computational technique for quantitatively evaluating instance explanations is to create a ground truth based on the input features that semantically contribute to the target class. For example, image segmentation maps (annotations of objects in images) are used to evaluate model generated saliency maps in weakly supervised object localization tasks [121]. Mohseni et al. [149] proposed a multi-layer *Human-Attention* baseline for feature-level evaluation of machine learning explanations. Their *Human-Attention* baseline provides a human-grounded feature attribution map with a higher level of granularity compared to object segmentation maps. Similarly, feature-level annotations have been used as the explanation ground truth in the text classification domain [53]. Other less accurate means of feature attribution like bounding box in images datasets have been used for quantitative evaluation of saliency maps. For instance, Du et al. [52] evaluated saliency maps generated from a CNN model by calculating pixel-wise IOU (intersection over union) of model-explanation bounding boxes and ground truth bounding boxes.

9.5 Role of User Interactions in XAI

Another important consideration in designing XAI systems is how to leverage user interactions to better support system understandability. The benefits of interactive system design have been previously explored in the topic of interactive machine learning [6, 7] for novice end-users. AI and data experts also benefit from interactive visual tools to improve model and task performance [57]. In this section, we discuss multiple examples of interaction design that support user understanding of the underlying black-box model.

Focusing on interactive design for AI-based systems for AI novices, Amershi et al. [6] reviewed multiple case studies that demonstrate the effectiveness of interactivity with a tight coupling between the algorithm and the user. They emphasize how interactive machine learning processes

allow the users to instantly examine the impact of their actions and adapt their next queries to improve outcomes. Such interactions allow users to test various inputs and learn about the model by creating *What-If* explanations [204]. Particularly, user-led cyles of trial and error help novices to understand how the machine learning model works and how to steer the model to improve results. In the context of XAI, Jongejan and Holbrook [29] present a study in which users draw images to see whether an image recognition algorithm can correctly recognize the intended sketch. Their system and study allows for interactive trial-and-error to explore how the algorithm works. In addition, their system provides example-based explanations in cases where the algorithm fails to correctly classify drawings. Another approach is to allow users to control or tune algorithmic parameters to achieve better results. For example, Kocielnik et al. [103] present a study in which users were able to freely control detection sensitivity in an AI assistant. Their results showed a significant effect on user perception of control and acceptance.

Visual analytics tools also support model understanding for expert users through interaction with algorithms. Examples including allowing data scientists and model experts to interactively explore model representations [86], analyze model training processes [129], and detect learning biases [27]. Also, embedded interaction techniques can support the exploration of very large deep learning networks. For instance, Hohman et al. [86] present multiple interactive features to select and filter of neurons and zoom and pan in feature representations to support AI experts in interpreting and reviewing trained models.

9.6 Generalization and Extension of the Framework

Our framework is extendable and compatible with existing AI-infused interface and interaction design guidelines. For example, Amershi et al. [7] propose 18 design guidelines for human-AI interaction design. Their guidelines are based on a review of a large number of AI-related design recommendation sources. They systematically validated guidelines through multiple rounds of evaluations with 49 design practitioners in 20 AI-infused products. Their design guidelines provide further details within the user interface design layer of our framework (Section 8.2) to guide the development of appropriate user interactions with model output and interactions. In other work, Dudley and Kristensson [54] present a review and characterization of user interface design principles for interactive machine learning systems. They propose a structural breakdown of interactive machine learning systems and present six principles to support system design. This work also benefits our framework by contributing practices of interactive machine learning design to the XAI system goals layer (Section 8.1) and the user interface design layer (Section 8.2) From the standpoint of evaluation methods, Mueller and Klein [153] discuss how common usability tests cannot address intelligent tools where software replicates human intelligence. They suggest new solutions are needed to allow the users to experience an AI-based tool's strengths and weaknesses. Likewise, our nested framework points out the potential for error propagation from the inner layers (e.g., interpretable algorithms design) to the outer layers (e.g., system outcomes) in the XAI system evaluation pole. The iterative back-and-forth between layers in the nested model encourages expert review of system outcomes, user-centered evaluation of the explainable interface, and computational evaluation of machine learning algorithms.

9.7 Limitations of the Framework

Our framework provides a basis for XAI system design in interdisciplinary teamwork and we have described our case study example to validate and improve the framework. The presented case study serves as a practical example of using our framework in a multidisciplinary collaborative XAI design and development effort. Our use case is the result of a year-long (and ongoing) research done by a team of eight university researchers with diverse backgrounds. The lessons learned and

pitfalls in our end-to-end implementation case study are incorporated in the presented design guidelines. However, no framework is perfect or entirely comprehensive. We acknowledge that the validity and usefulness of a framework are to be proven in practice with further case studies. In our future work, we plan to run multiple validation case studies to examine practicality and usefulness of this framework.

Moreover, this framework has a common limitation of many multidisciplinary design frameworks of being light on specific details at each step. Rather than contributing detailed guidelines for each framework layer, the framework is intended to pave the path for efficient collaboration among and within different teams, which is essential for XAI system design given the inherently interdisciplinary nature of this field. This higher level of freedom allows for extendability with other design guidelines (see the discussion in Section 9.6) to integrate with more tailored approaches for specific domains. Additionally, the diversity of design goals and evaluation methods at each layer can help maintain the balance of attention from the design team to different aspects of XAI system.

10 CONCLUSION

We reviewed XAI-related research to organize multiple XAI design goals and evaluation measures. Table 2 presents our categorization of selected existing design and evaluation methods that organizes literature along three perspectives: *design goals*, *evaluation methods*, and the *targeted users* of the XAI system. We provide summarized ready-to-use tables of evaluation methods and recommendations for different goals in XAI research. Our categorization revealed the necessity of an interdisciplinary effort for designing and evaluating XAI systems. We want to draw attention to related resources in the social sciences that can facilitate the extent of social and cognitive aspects of explanations. To address these issues, we proposed a design and evaluation framework that connects design goals and evaluation methods for end-to-end XAI systems design, as presented through a model and a series of guidelines. We hope our framework drives further discussion about the interplay between design and evaluation of explainable artificial intelligent systems. Although the presented framework is organized to provide a high-level guideline for a multidisciplinary effort to build XAI systems, it is not meant to offer all aspects of interface and interaction design and development of interpretable machine learning techniques. Lastly, we briefly discussed additional considerations for XAI designers to benefit from the body of knowledge of XAI system design and evaluation.

11 ACKNOWLEDGMENTS

The authors would like to thank anonymous reviewers for their helpful comments on earlier versions of this manuscript. The work in this paper is supported by the DARPA XAI program under N66001-17-2-4031 and by NSF award 1900767. The views and conclusions in this paper are those of the authors and should not be interpreted as representing any funding agencies.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 582.

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[3] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*. 9505–9515.

[4] Yongsu Ahn and Yu-Ru Lin. 2019. Fairsight: visual analytics for fairness in decision making. *IEEE Transactions on Visualization and Computer Graphics* (2019).

[5] Eric Alexander and Michael Gleicher. 2015. Task-driven comparison of topic models. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2015), 320–329.

[6] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120.

[7] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 3.

[8] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).

[9] Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* 20, 3 (2018), 973–989.

[10] Stavros Antifakos, Nicky Kern, Bernt Schiele, and Adrian Schwaninger. 2005. Towards improving trust in context-aware systems by displaying system confidence. In *Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Devices & Services*. ACM, 9–14.

[11] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[12] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. 2014. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755* (2014).

[13] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10, 7 (2015), e0130140.

[14] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. 2010. How to explain individual classification decisions. *Journal of Machine Learning Research* 11, Jun (2010), 1803–1831.

[15] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.

[16] Victoria Bellotti and Keith Edwards. 2001. Intelligibility and accountability: human considerations in context-aware systems. *Human–Computer Interaction* 16, 2-4 (2001), 193–212.

[17] Shlomo Berkovsky, Ronnie Taib, and Dan Conway. 2017. How to Recommend?: User Trust Factors in Movie Recommender Systems. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces (IUI '17)*. ACM, New York, NY, USA, 287–300. https://doi.org/10.1145/3025171.3025209

[18] Daniel M Best, Alex Endert, and Daniel Kidwell. 2014. 7 key challenges for visualization in cyber network defense. In *Proceedings of the Eleventh Workshop on Visualization for Cyber Security*. ACM, 33–40.

[19] Mustafa Bilgic and Raymond J Mooney. 2005. Explaining recommendations: Satisfaction vs. promotion. In *Beyond Personalization Workshop, IUI*, Vol. 5. 153.

[20] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. "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*. ACM, 377.

[21] Philip Bobko, Alex J Barelka, and Leanne M Hirshfield. 2014. The construct of state-level suspicion: A model and research agenda for automated and information technology (IT) contexts. *Human Factors* 56, 3 (2014), 489–508.

[22] Mariusz Bojarski, Anna Choromanska, Krzysztof Choromanski, Bernhard Firner, Larry J Ackel, Urs Muller, Phil Yeres, and Karol Zieba. 2018. Visualbackprop: Efficient visualization of cnns for autonomous driving. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1–8.

[23] Engin Bozdag and Jeroen van den Hoven. 2015. Breaking the filter bubble: democracy and design. *Ethics and Information Technology* 17, 4 (2015), 249–265.

[24] Nicholas Bryan and Gautham Mysore. 2013. An efficient posterior regularized latent variable model for interactive sound source separation. In *International Conference on Machine Learning*. 208–216.

[25] Andrea Bunt, Matthew Lount, and Catherine Lauzon. 2012. 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*. ACM, 169–178.

[26] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *International Conference on Healthcare Informatics (ICHI)*. IEEE, 160–169.

[27] Angel Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. 2019. FairVis: visual analytics for discovering intersectional bias in machine learning. *IEEE Conference on Visual Analytics Science and Technology (VAST)* (2019).

[28] Béatrice Cahour and Jean-François Forzy. 2009. Does projection into use improve trust and exploration? An example with a cruise control system. *Safety Science* 47, 9 (2009), 1260–1270.

[29] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 258–262.

[30] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[31] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1721–1730.

[32] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 339–348.

[33] Jaegul Choo, Hanseung Lee, Jaeyeon Kihm, and Haesun Park. 2010. iVisClassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*. IEEE, 27–34.

[34] Jaegul Choo and Shixia Liu. 2018. Visual analytics for explainable deep learning. *IEEE Computer Graphics and Applications* 38, 4 (2018), 84–92.

[35] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[36] Michael Chromik, Malin Eiband, Sarah Theres Völkel, and Daniel Buschek. 2019. Dark patterns of explainability, transparency, and user control for intelligent systems.. In *IUI Workshops*.

[37] Lingyang Chu, Xia Hu, Juhua Hu, Lanjun Wang, and Jian Pei. 2018. Exact and consistent interpretation for piecewise linear neural networks: A closed form solution. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1244–1253.

[38] Miruna-Adriana Clinciu and Helen Hastie. 2019. A survey of explainable AI terminology. In *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*. 8–13.

[39] Sven Coppers, Jan Van den Bergh, Kris Luyten, Karin Coninx, Iulianna Van der Lek-Ciudin, Tom Vanallemeersch, and Vincent Vandeghinste. 2018. Intellingo: An intelligible translation environment. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 524.

[40] Enrico Costanza, Joel E Fischer, James A Colley, Tom Rodden, Sarvapali D Ramchurn, and Nicholas R Jennings. 2014. Doing the laundry with agents: a field trial of a future smart energy system in the home. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 813–822.

[41] William Curran, Travis Moore, Todd Kulesza, Weng-Keen Wong, Sinisa Todorovic, Simone Stumpf, Rachel White, and Margaret Burnett. 2012. Towards recognizing cool: can end users help computer vision recognize subjective attributes of objects in images?. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*. ACM, 285–288.

[42] Abhishek Das, Harsh Agrawal, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2016. Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions?. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://computing.ece.vt.edu/~abhshkdz/vqa-hat/

[43] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding* 163 (2017), 90–100.

[44] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies* 2015, 1 (2015), 92–112.

[45] Nicholas Diakopoulos. 2014. Algorithmic-Accountability: the investigation of Black Boxes. *Tow Center for Digital Journalism* (2014).

[46] Nicholas Diakopoulos. 2017. Enabling Accountability of Algorithmic Media: Transparency as a Constructive and Critical Lens. In *Transparent Data Mining for Big and Small Data*. Springer, 25–43.

[47] Jonathan Dodge, Sean Penney, Andrew Anderson, and Margaret M Burnett. 2018. What Should Be in an XAI Explanation? What IFT Reveals.. In *IUI Workshops*.

[48] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[49] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Christopher Bavitz, Samuel J Gershman, David O'Brien, Stuart Shieber, Jim Waldo, David Weinberger, and Alexandra Wood. 2017. Accountability of AI Under the Law: The Role of Explanation. *Berkman Center Research Publication Forthcoming* (2017), 18–07.

[50] James K Doyle, Michael J Radzicki, and W Scott Trees. 2008. Measuring change in mental models of complex dynamic systems. In *Complex Decision Making*. Springer, 269–294.

[51] Fan Du, Catherine Plaisant, Neil Spring, Kenyon Crowley, and Ben Shneiderman. 2019. EventAction: A Visual Analytics Approach to Explainable Recommendation for Event Sequences. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 9, 4 (2019), 1–31.

[52] Mengnan Du, Ninghao Liu, Qingquan Song, and Xia Hu. 2018. Towards explanation of dnn-based prediction with guided feature inversion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1358–1367.

[53] M. Du, N. Liu, F. Yang, and X. Hu. 2019. Learning credible deep neural networks with rationale regularization. In *2019 IEEE International Conference on Data Mining (ICDM)*. 150–159.

[54] John J Dudley and Per Ola Kristensson. 2018. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 8.

[55] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. 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, LBW0243.

[56] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing transparency design into practice. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. ACM, New York, NY, USA, 211–223. https://doi.org/10.1145/3172944.3172961

[57] A Endert, W Ribarsky, C Turkay, BL Wong, Ian Nabney, I Díaz Blanco, and F Rossi. 2017. The state of the art in integrating machine learning into visual analytics. In *Computer Graphics Forum*, Vol. 36. Wiley Online Library, 458–486.

[58] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. I always assumed that I wasn't really that close to [her]: Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 153–162.

[59] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. "Be careful; things can be worse than they appear": Understanding Biased Algorithms and Users' Behavior around Them in Rating Platforms. In *Eleventh International AAAI Conference on Web and Social Media*.

[60] Raquel Florez-Lopez and Juan Manuel Ramon-Jeronimo. 2015. Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. A correlated-adjusted decision forest proposal. *Expert Systems with Applications* 42, 13 (2015), 5737–5753.

[61] Ruth C Fong and Andrea Vedaldi. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*. 3429–3437.

[62] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How should I explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies* 72, 4 (2014), 367–382.

[63] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. 2019. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*. 9273–9282.

[64] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 80–89.

[65] Alyssa Glass, Deborah L McGuinness, and Michael Wolverton. 2008. Toward establishing trust in adaptive agents. In *Proceedings of the 13th International Conference on Intelligent User Interfaces*. ACM, 227–236.

[66] John Goodall, Eric D Ragan, Chad A Steed, Joel W Reed, G David Richardson, Kelly MT Huffer, Robert A Bridges, and Jason A Laska. 2018. Situ: Identifying and Explaining Suspicious Behavior in Networks. *IEEE Transactions on Visualization and Computer Graphics* (2018).

[67] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine* 38, 3 (2017), 50–57.

[68] Colin M Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L Toombs. 2018. The dark (patterns) side of UX design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 534.

[69] Shirley Gregor and Izak Benbasat. 1999. Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice. *Management Information Systems Quarterly* 23, 4 (1999), 2.

[70] Alex Groce, Todd Kulesza, Chaoqiang Zhang, Shalini Shamasunder, Margaret Burnett, Weng-Keen Wong, Simone Stumpf, Shubhomoy Das, Amber Shinsel, Forrest Bice, et al. 2014. You are the only possible oracle: Effective test selection for end users of interactive machine learning systems. *IEEE Transactions on Software Engineering* 40, 3 (2014), 307–323.

[71] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)* 51, 5 (2018), 93.

[72] David Gunning. 2017. Explainable artificial intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA)* (2017).

[73] Aniko Hannak, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. 2013. Measuring personalization of web search. In *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 527–538.
[74] Steven R Haynes, Mark A Cohen, and Frank E Ritter. 2009. Designs for explaining intelligent agents. *International Journal of Human-Computer Studies* 67, 1 (2009), 90–110.
[75] Jeffrey Heer. 2019. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1844–1850.
[76] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*. Springer, 793–811.
[77] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*. ACM, 241–250.
[78] Bernease Herman. 2017. The Promise and Peril of Human Evaluation for Model Interpretability. *arXiv preprint arXiv:1711.07414* (2017).
[79] Robert Hoffman, Tim Miller, Shane T Mueller, Gary Klein, and William J Clancey. 2018. Explaining explanation, part 4: a deep dive on deep nets. *IEEE Intelligent Systems* 33, 3 (2018), 87–95.
[80] Robert R Hoffman. 2017. Theory concepts measures but policies metrics. In *Macrocognition Metrics and Scenarios*. CRC Press, 35–42.
[81] Robert R Hoffman, John K Hawley, and Jeffrey M Bradshaw. 2014. Myths of automation, part 2: Some very human consequences. *IEEE Intelligent Systems* 29, 2 (2014), 82–85.
[82] Robert R Hoffman, Matthew Johnson, Jeffrey M Bradshaw, and Al Underbrink. 2013. Trust in automation. *IEEE Intelligent Systems* 28, 1 (2013), 84–88.
[83] Robert R Hoffman and Gary Klein. 2017. Explaining explanation, part 1: theoretical foundations. *IEEE Intelligent Systems* 32, 3 (2017), 68–73.
[84] Robert R Hoffman, Shane T Mueller, and Gary Klein. 2017. Explaining explanation, part 2: empirical foundations. *IEEE Intelligent Systems* 32, 4 (2017), 78–86.
[85] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).
[86] Fred Hohman, Haekyu Park, Caleb Robinson, and Duen Horng Polo Chau. 2019. Summit: scaling deep learning interpretability by visualizing activation and attribution summarizations. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 1096–1106.
[87] Fred Hohman, Arjun Srinivasan, and Steven M. Drucker. 2019. TeleGam: combining visualization and verbalization for interpretable machine learning. *IEEE Visualization Conference (VIS)* (2019).
[88] Fred Matthew Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. 2018. Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Transactions on Visualization and Computer Graphics* (2018).
[89] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. 2016. User trust in intelligent systems: A journey over time. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. ACM, 164–168.
[90] Kristina Höök. 2000. Steps to take before intelligent user interfaces become real. *Interacting with Computers* 12, 4 (2000), 409–426.
[91] Philip N Howard and Bence Kollanyi. 2016. Bots, #StrongerIn, and #Brexit: computational propaganda during the UK-EU referendum. (2016).
[92] Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2014. Interactive topic modeling. *Machine Learning* 95, 3 (2014), 423–469.
[93] Shagun Jhaver, Yoni Karpfen, and Judd Antin. 2018. Algorithmic anxiety and coping strategies of Airbnb hosts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 421.
[94] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71.
[95] Minsuk Kahng, Pierre Y Andrews, Aditya Kalro, and Duen Horng Polo Chau. 2018. ActiVis: visual exploration of industry-scale deep neural network models. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 88–97.
[96] Matthew Kay, Tara Kola, Jessica R Hullman, and Sean A Munson. 2016. When (ish) is my bus?: User-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5092–5103.
[97] Frank C Keil. 2006. Explanation and understanding. *Annu. Rev. Psychol.* 57 (2006), 227–254.
[98] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems*. 2280–2288.

[99] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *International Conference on Machine Learning*. 2673–2682.

[100] Jaedeok Kim and Jingoo Seo. 2017. Human understandable explanation extraction for black-box classification models based on matrix factorization. *arXiv preprint arXiv:1709.06201* (2017).

[101] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2019. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 267–280.

[102] Gary Klein. 2018. Explaining explanation, part 3: The causal landscape. *IEEE Intelligent Systems* 33, 2 (2018), 83–88.

[103] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[104] Johannes Kraus, David Scholz, Dina Stiegemeier, and Martin Baumann. 2019. The more you know: Trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency. *Human Factors* (2019), 0018720819853686.

[105] Josua Krause, Aritra Dasgupta, Jordan Swartz, Yindalon Aphinyanaphongs, and Enrico Bertini. 2017. A workflow for visual diagnostics of binary classifiers using instance-level explanations. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 162–172.

[106] Josua Krause, Adam Perer, and Enrico Bertini. 2014. INFUSE: interactive feature selection for predictive modeling of high dimensional data. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 1614–1623.

[107] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5686–5697.

[108] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, 126–137.

[109] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell Me More?: The Effects of Mental Model Soundness on Personalizing an Intelligent Agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 1–10.

[110] Todd Kulesza, Simone Stumpf, Margaret Burnett, Weng-Keen Wong, Yann Riche, Travis Moore, Ian Oberst, Amber Shinsel, and Kevin McIntosh. 2010. Explanatory debugging: Supporting end-user debugging of machine-learned programs. In *Visual Languages and Human-Centric Computing (VL/HCC), 2010 IEEE Symposium on*. IEEE, 41–48.

[111] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *Visual Languages and Human-Centric Computing (VL/HCC), 2013 IEEE Symposium on*. IEEE, 3–10.

[112] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. 2019. Human evaluation of models built for interpretability. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 59–67.

[113] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. 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*. ACM, 1675–1684.

[114] Ellen J Langer, Arthur Blank, and Benzion Chanowitz. 1978. The mindlessness of ostensibly thoughtful action: The role of "placebic" information in interpersonal interaction. *Journal of Personality and Social Psychology* 36, 6 (1978), 635.

[115] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 182.

[116] Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. Working with machines: The impact of algorithmic and data-driven management on human workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1603–1612.

[117] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2017. Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philosophy & Technology* (2017), 1–17.

[118] Piyawat Lertvittayakumjorn and Francesca Toni. 2019. Human-grounded Evaluations of Explanation Methods for Text Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5198–5208.

[119] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, David Madigan, et al. 2015. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics* 9, 3 (2015), 1350–1371.

[120] Alexander Lex, Marc Streit, H-J Schulz, Christian Partl, Dieter Schmalstieg, Peter J Park, and Nils Gehlenborg. 2012. StratomeX: Visual Analysis of Large-Scale Heterogeneous Genomics Data for Cancer Subtype Characterization. In *Computer Graphics Forum*, Vol. 31. Wiley Online Library, 1175–1184.

[121] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. 2018. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9215–9223.

[122] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Attention bridging network for knowledge transfer. In *Proceedings of the IEEE International Conference on Computer Vision*. 5198–5207.

[123] Brian Lim. 2011. Improving Understanding, Trust, and Control with Intelligibility in Context-Aware Applications. *Human-Computer Interaction* (2011).

[124] Brian Y Lim and Anind K Dey. 2009. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th International Conference on Ubiquitous Computing*. ACM, 195–204.

[125] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. 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*. ACM, 2119–2128.

[126] Brian Y Lim, Qian Yang, Ashraf M Abdul, and Danding Wang. 2019. Why these explanations? Selecting intelligibility types for explanation Goals. In *IUI Workshops*.

[127] Zachary C Lipton. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490* (2016).

[128] Mengchen Liu, Shixia Liu, Xizhou Zhu, Qinying Liao, Furu Wei, and Shimei Pan. 2016. An uncertainty-aware approach for exploratory microblog retrieval. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 250–259.

[129] Mengchen Liu, Jiaxin Shi, Kelei Cao, Jun Zhu, and Shixia Liu. 2018. Analyzing the training processes of deep generative models. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 77–87.

[130] Mengchen Liu, Jiaxin Shi, Zhen Li, Chongxuan Li, Jun Zhu, and Shixia Liu. 2017. Towards better analysis of deep convolutional neural networks. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 91–100.

[131] Shixia Liu, Xiting Wang, Jianfei Chen, Jim Zhu, and Baining Guo. 2014. TopicPanorama: A full picture of relevant topics. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*. IEEE, 183–192.

[132] Tania Lombrozo. 2006. The structure and function of explanations. *Trends in Cognitive Sciences* 10, 10 (2006), 464–470.

[133] Tania Lombrozo. 2009. Explanation and categorization: How "why?" informs "what?". *Cognition* 110, 2 (2009), 248–253.

[134] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*. 4765–4774.

[135] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, Nov (2008), 2579–2605.

[136] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *11th Australasian Conference on Information Systems*, Vol. 53. Citeseer, 6–8.

[137] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).

[138] Sarah Mennicken, Jo Vermeulen, and Elaine M Huang. 2014. From today's augmented houses to tomorrow's smart homes: new directions for home automation research. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 105–115.

[139] Stephanie M Merritt, Heather Heimbaugh, Jennifer LaChapell, and Deborah Lee. 2013. I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors* 55, 3 (2013), 520–534.

[140] Miriah Meyer, Michael Sedlmair, P Samuel Quinan, and Tamara Munzner. 2015. The nested blocks and guidelines model. *Information Visualization* 14, 3 (2015), 234–249.

[141] Debra Meyerson, Karl E Weick, and Roderick M Kramer. 1996. Swift trust and temporary groups. *Trust in organizations: Frontiers of theory and research* 166 (1996), 195.

[142] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[143] Yao Ming, Shaozu Cao, Ruixiang Zhang, Zhen Li, Yuanzhe Chen, Yangqiu Song, and Huamin Qu. 2017. Understanding hidden memories of recurrent neural networks. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 13–24.

[144] Yao Ming, Huamin Qu, and Enrico Bertini. 2018. Rulematrix: Visualizing and understanding classifiers with rules. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2018), 342–352.

[145] Brent Mittelstadt. 2016. Automation, algorithms, and politics: Auditing for transparency in content personalization systems. *International Journal of Communication* 10 (2016), 12.

[146] Sina Mohseni, Akshay Jagadeesh, and Zhangyang Wang. 2019. Predicting model failure using saliency maps in autonomous driving systems. *ICML Workshop on Uncertainty & Robustness in Deep Learning* (2019).

[147] Sina Mohseni, Mandar Pitale, Vasu Singh, and Zhangyang Wang. 2020. Practical solutions for machine learning safety in autonomous vehicles. In *The AAAI Workshop on Artificial Intelligence Safety (Safe AI)*.

[148] Sina Mohseni, Eric Ragan, and Xia Hu. 2019. Open issues in combating fake news: Interpretability as an opportunity. *arXiv preprint arXiv:1904.03016* (2019).

[149] Sina Mohseni and Eric D Ragan. 2018. A human-grounded evaluation benchmark for local explanations of machine learning. *arXiv preprint arXiv:1801.05075* (2018).

[150] Sina Mohseni, Fan Yang, Shiva Pentyala, Mengnan Du, Yi Liu, Nic Lupfer, Xia Hu, Shuiwang Ji, and Eric Ragan. 2020. Trust evolution over time in explainable AI for fake news detection. *Fair & Responsible AI Workshop at CHI 2020* (2020).

[151] Christoph Molnar. 2019. *Interpretable machine learning*. Lulu. com.

[152] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2017. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* (2017).

[153] Shane T Mueller and Gary Klein. 2011. Improving users' mental models of intelligent software tools. *IEEE Intelligent Systems* 26, 2 (2011), 77–83.

[154] Bonnie M Muir. 1987. Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies* 27, 5-6 (1987), 527–539.

[155] Tamara Munzner. 2009. A nested process model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics* 6 (2009), 921–928.

[156] Brad A Myers, David A Weitzman, Andrew J Ko, and Duen H Chau. 2006. Answering why and why not questions in user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 397–406.

[157] Andrew P Norton and Yanjun Qi. 2017. Adversarial-playground: A visualization suite showing how adversarial examples fool deep learning. In *Visualization for Cyber Security (VizSec), 2017 IEEE Symposium on*. IEEE, 1–4.

[158] Florian Nothdurft, Felix Richter, and Wolfgang Minker. 2014. Probabilistic human-computer trust handling. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. 51–59.

[159] Mahsan Nourani, Dondald Honeycutt, Jeremy Block, Rahman Tahrima Roy, Chiradeep, Eric D. Ragan, and Vibhav Gogate. 2020. Investigating the importance of first impressions and explainable AI with interactive video analysis. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.

[160] Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D Ragan. 2019. 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*, Vol. 7. 97–105.

[161] Besmira Nushi, Ece Kamar, and Eric Horvitz. 2018. Towards accountable AI: Hybrid human-machine analyses for characterizing system failure. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.

[162] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. 2018. The Building Blocks of Interpretability. *Distill* (2018). https://doi.org/10.23915/distill.00010 https://distill.pub/2018/building-blocks.

[163] Cathy O'Neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

[164] Sean Penney, Jonathan Dodge, Claudia Hilderbrand, Andrew Anderson, Logan Simpson, and Margaret Burnett. 2018. Toward foraging for understanding of StarCraft agents: An empirical study. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. ACM, New York, NY, USA, 225–237. https://doi.org/10.1145/3172944.3172946

[165] Nicola Pezzotti, Thomas Höllt, Jan Van Gemert, Boudewijn PF Lelieveldt, Elmar Eisemann, and Anna Vilanova. 2018. DeepEyes: Progressive visual analytics for designing deep neural networks. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 98–108.

[166] Nina Poerner, Hinrich Schütze, and Benjamin Roth. 2018. Evaluating neural network explanation methods using hybrid documents and morphological prediction. In *56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

[167] Brett Poulin, Roman Eisner, Duane Szafron, Paul Lu, Russell Greiner, David S Wishart, Alona Fyshe, Brandon Pearcy, Cam MacDonell, and John Anvik. 2006. Visual explanation of evidence with additive classifiers. In *Proceedings Of The National Conference On Artificial Intelligence*, Vol. 21. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 1822.

[168] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810* (2018).

[169] Pearl Pu and Li Chen. 2006. Trust building with explanation interfaces. In *Proceedings of the 11th International Conference on Intelligent User Interfaces*. ACM, 93–100.

[170] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 103.

[171] Emilee Rader and Rebecca Gray. 2015. 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*. ACM, 173–182.
[172] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1135–1144.
[173] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence*.
[174] Caleb Robinson, Fred Hohman, and Bistra Dilkina. 2017. A deep learning approach for population estimation from satellite imagery. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*. ACM, 47–54.
[175] Marko Robnik-Šikonja and Marko Bohanec. 2018. Perturbation-based explanations of prediction models. In *Human and Machine Learning*. Springer, 159–175.
[176] Stephanie Rosenthal, Sai P Selvaraj, and Manuela Veloso. 2016. Verbalization: narration of autonomous robot experience. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. 862–868.
[177] Andrew Slavin Ross and Finale Doshi-Velez. 2018. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Thirty-second AAAI Conference on Artificial Intelligence*.
[178] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 2662–2670. https://doi.org/10.24963/ijcai.2017/371
[179] Stephen Rudolph, Anya Savikhin, and David S Ebert. 2009. Finvis: Applied visual analytics for personal financial planning. In *Visual Analytics Science and Technology, 2009. IEEE Symposium on*. Citeseer, 195–202.
[180] Dominik Sacha, Michael Sedlmair, Leishi Zhang, John Aldo Lee, Daniel Weiskopf, Stephen North, and Daniel Keim. 2016. Human-centered machine learning through interactive visualization. In *24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. 641–646.
[181] Dominik Sacha, Hansi Senaratne, Bum Chul Kwon, Geoffrey Ellis, and Daniel A Keim. 2016. The role of uncertainty, awareness, and trust in visual analytics. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 240–249.
[182] Bahador Saket, Arjun Srinivasan, Eric D Ragan, and Alex Endert. 2017. Evaluating interactive graphical encodings for data visualization. *IEEE Transactions on Visualization and Computer Graphics* 24, 3 (2017), 1316–1330.
[183] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. 2017. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems* 28, 11 (2017), 2660–2673.
[184] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns Into Productive Inquiry* (2014), 1–23.
[185] Martin Schaffernicht and Stefan N Groesser. 2011. A comprehensive method for comparing mental models of dynamic systems. *European Journal of Operational Research* 210, 1 (2011), 57–67.
[186] Ute Schmid, Christina Zeller, Tarek Besold, Alireza Tamaddoni-Nezhad, and Stephen Muggleton. 2016. How does predicate invention affect human comprehensibility?. In *International Conference on Inductive Logic Programming*. Springer, 52–67.
[187] Philipp Schmidt and Felix Biessmann. 2019. Quantifying interpretability and trust in machine learning systems. *arXiv preprint arXiv:1901.08558* (2019).
[188] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.
[189] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 3145–3153.
[190] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).
[191] Daniel Smilkov, Shan Carter, D Sculley, Fernanda B Viégas, and Martin Wattenberg. 2017. Direct-manipulation visualization of deep networks. *arXiv preprint arXiv:1708.03788* (2017).
[192] Thilo Spinner, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady. 2019. explAIner: A visual analytics framework for interactive and explainable machine learning. *IEEE Transactions on Visualization and Computer Graphics* (2019).
[193] Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister, and Alexander M Rush. 2018. LstmVis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 667–676.

[194] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies* 67, 8 (2009), 639–662.

[195] Simone Stumpf, Simonas Skrebe, Graeme Aymer, and Julie Hobson. 2018. Explaining smart heating systems to discourage fiddling with optimized behavior. (2018).

[196] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Commun. ACM* 56, 5 (2013), 44–54.

[197] Jiliang Tang, Huiji Gao, Huan Liu, and Atish Das Sarma. 2012. eTrust: Understanding trust evolution in an online world. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 253–261.

[198] Christina Tikkinen-Piri, Anna Rohunen, and Jouni Markkula. 2018. EU general data protection regulation: Changes and implications for personal data collecting companies. *Computer Law & Security Review* 34, 1 (2018), 134–153.

[199] Nava Tintarev and Judith Masthoff. 2011. Designing and evaluating explanations for recommender systems. In *Recommender Systems Handbook*. Springer, 479–510.

[200] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. 2018. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552* (2018).

[201] Matteo Turilli and Luciano Floridi. 2009. The ethics of information transparency. *Ethics and Information Technology* 11, 2 (2009), 105–112.

[202] Jo Vermeulen, Geert Vanderhulst, Kris Luyten, and Karin Coninx. 2010. PervasiveCrystal: Asking and answering why and why not questions about pervasive computing applications. In *Intelligent Environments (IE), 2010 Sixth International Conference on*. IEEE, 271–276.

[203] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology* 31 (2017), 841.

[204] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 601, 15 pages.

[205] Fulton Wang and Cynthia Rudin. 2015. Falling rule lists. In *Artificial Intelligence and Statistics*. 1013–1022.

[206] Qianwen Wang, Jun Yuan, Shuxin Chen, Hang Su, Huamin Qu, and Shixia Liu. 2019. Visual genealogy of deep neural networks. *IEEE Transactions on Visualization and Computer Graphics* (2019).

[207] Daniel S. Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. *Commun. ACM* 62, 6 (May 2019), 70âĂŞ79.

[208] Adrian Weller. 2017. Challenges for transparency. *arXiv preprint arXiv:1708.01870* (2017).

[209] Gesa Wiegand, Matthias Schmidmaier, Thomas Weber, Yuanting Liu, and Heinrich Hussmann. 2019. I drive-you trust: Explaining driving behavior of autonomous cars. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, LBW0163.

[210] James A Wise, James J Thomas, Kelly Pennock, David Lantrip, Marc Pottier, Anne Schur, and Vern Crow. 1995. Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In *Information Visualization, 1995. Proceedings*. IEEE, 51–58.

[211] Kanit Wongsuphasawat, Daniel Smilkov, James Wexler, Jimbo Wilson, Dandelion Mane, Doug Fritz, Dilip Krishnan, Fernanda B Viégas, and Martin Wattenberg. 2017. Visualizing dataflow graphs of deep learning models in tensorflow. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2017), 1–12.

[212] Samuel C Woolley. 2016. Automating power: Social bot interference in global politics. *First Monday* 21, 4 (2016).

[213] Mike Wu, Michael C Hughes, Sonali Parbhoo, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez. 2018. Beyond sparsity: Tree regularization of deep models for interpretability. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[214] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[215] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding neural networks through deep visualization. *In ICML Deep Learning Workshop 2015* (2015).

[216] Rulei Yu and Lei Shi. 2018. A user-based taxonomy for deep learning visualization. *Visual Informatics* 2, 3 (2018), 147–154.

[217] Tom Zahavy, Nir Ben-Zrihem, and Shie Mannor. 2016. Graying the black box: Understanding DQNs. In *International Conference on Machine Learning*. 1899–1908.

[218] Tal Zarsky. 2016. The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values* 41, 1 (2016), 118–132.

[219] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*. Springer, 818–833.

[220] Quanshi Zhang, Wenguan Wang, and Song-Chun Zhu. 2018. Examining cnn representations with respect to dataset bias. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[221] Quan-shi Zhang and Song-Chun Zhu. 2018. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering* 19, 1 (2018), 27–39.

[222] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. 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 (FAT* '20)*.

[223] Zijian Zhang, Jaspreet Singh, Ujwal Gadiraju, and Avishek Anand. 2019. Dissonance between human and machine understanding. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 56.

[224] Wen Zhong, Cong Xie, Yuan Zhong, Yang Wang, Wei Xu, Shenghui Cheng, and Klaus Mueller. 2017. Evolutionary visual analysis of deep neural networks. In *ICML Workshop on Visualization for Deep Learning*.

[225] Jichen Zhu, Antonios Liapis, Sebastian Risi, Rafael Bidarra, and G Michael Youngblood. 2018. Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 1–8.

[226] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. 2017. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595* (2017).