A Survey of Evaluation Methods and Measures for Interpretable Machine Learning

Sina Mohseni
Texas A&M University
College Station, USA
sina.mohseni@tamu.edu

Niloofar Zarei
Texas A&M University
College Station, USA
n.zarei.3001@tamu.edu

Eric D. Ragan
University of Florida
Gainesville, USA
eragan@ufl.edu

ABSTRACT
The need for interpretable and accountable intelligent system gets sensible as artificial intelligence plays more role in human life. Explainable artificial intelligence systems can be a solution by self-explaining the reasoning behind the decisions and predictions of the intelligent system. Researchers from different disciplines work together to define, design and evaluate interpretable intelligent systems for the user. Our work supports the different evaluation goals in interpretable machine learning research by a thorough review of evaluation methodologies used in machine-explanation research across the fields of human-computer interaction, visual analytics, and machine learning. We present a 2D categorization of interpretable machine learning evaluation methods and show a mapping between user groups and evaluation measures. Further, we address the essential factors and steps for a right evaluation plan by proposing a nested model for design and evaluation of explainable artificial intelligence systems.

INTRODUCTION
Although the fantasy of the "super-human" may still be long way off, the "super-AI" is much closer our current reality. Tech giants like Google, Facebook, and Amazon, have already 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 [127, 52] were yet another sign of how susceptible our lives are to the power of artificial intelligence (AI) and big data [92]. 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 trustable AI grows as we get closer to inventing the unimaginable super-AI.

Explainable artificial intelligence (XAI) systems can be a solution to predictable and accountable AI by explaining AI decision-making processes and logic for end users [41]. One can define XAI 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 driven much attention from multiple different communities, interpretability of intelligent systems has also been studied in numerous contexts [96, 38]. The study of personalized agents, recommendation systems, and critical decision-making tasks (e.g., medical analysis, powergrid control) span application domains interested in machine-learning explanation and AI transparency. The legal right to explanations has also 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 requirement from AI systems [37].

Obviously, finding answers to such broad definitions of XAI requires multidisciplinary research efforts. 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. To study the concerns of end-users, research in human-computer interaction (HCI) focuses on end-user needs such as user trust and understanding of machine generate explanations. Psychology research also studies the fundamentals of human understanding, interpretability, the structure of explanations.

Looking at the broad spectrum of research on XAI, it is clear that scholars from different disciplines have different goals in mind. In other words, even though XAI research are following the general goals of AI interpretability, different measures and metrics are used to evaluate the XAI goals. For example, numerical methods are practiced in machine learning field to evaluate computational interpretability, while human interpretability and human-subjects evaluations are often the primary goals in HCI and visual analytics communities. In this regard, although there seems to be a mismatch in objectives for the scholars to define, design, and evaluate the concept of XAI, there is no doubt that certain types of convergence in XAI research are necessary to achieve the benefits of XAI.

This paper presents a survey intended to share knowledge and experiences of machine-explanation evaluation methods across multiple disciplines. To support the diverse evaluation goals 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 evaluation methods and show a mapping between user groups and evaluation measures. We further address the essential factors and steps for an effective evaluation plan by proposing a nested model for design and evaluation of explainable artificial intelligence systems. The main contribution of this model is to give guidance on what evaluation measures are appropriate to use at which stages of study considering the XAI end-users.

BACKGROUND

Nowadays, algorithms analyze user data and affect the decision-making process for millions of people on matters like employment, insurance rates, loan rates, and even criminal justice [21]. However, these algorithms that serve critical roles in many industries have their own disadvantages that can result in discrimination [24, 118] and unfair decision-making [92]. For instance, recently, news feed and targeted advertising algorithms in social media have taken much attention for leading the lack of information diversity in social media [13]. A significant part of trouble could be because algorithmic decision-making systems—unlike recommender systems—do not let users choose between the recommended items and instead pick the most relevant content or option for the user.

Bellotti and Keith [7] 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 which is hidden to the end users and causing blind faith [130]. Key benefits of algorithmic transparency and interpretability include user awareness [4], bias and discrimination detection [26, 118], interpretable behavior of intelligent systems [70], and accountability for users [27]. Furthermore, considering the increasing examples of discrimination and other lawful 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 [30, 84, 123].

Auditing Inexplicable AI

Researchers audit algorithms to study bias and discrimination in algorithmic decision making [111] and study user awareness about the effect of these algorithms [32]. Auditing algorithms is a mechanism of investigating algorithms' functionality to detect bias and other unwanted algorithm behaviors from the outside without the need to know about specific design details. Auditing methods focus on problematic effects in algorithmic decision making and take technical implementation details aside. 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 [24] and crowdsourcing [42] 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 work by Eslami et al. [33], in which they analyzed user reviews in three hotel booking websites to study user awareness of bias in online rating algorithms. These examples show that auditing algorithms is a valuable but time-taking process that could not be scaled easily to high number of algorithms. This calls for new research for solutions toward algorithmic transparency.

Explainable AI

Along with the aforementioned methods for supporting transparency, machine learning explanations also became a common approach to achieve transparency in many applications like social media, e-commerce and data-driven management of human workers [119, 120, 68]. The systems in these contexts generate explanations and describe the reasoning behind machine-learning decisions and predictions. Machine-learning explanations enable users to understand how the data is processed and supports awareness of possible bias and systems malfunctions. For example, to measure users perception of justice in intelligent decision making, Binns et al. [11] studied explanations in daily-life systems 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 system. In similar work studying knowledge of social media algorithms, Radar et al. [98] ran a crowdsourced study to see how different types of explanations effects users' beliefs on newsfeed algorithm 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 behavior. Keeping human-users accountable is another crucial aspect in transparent machine learning. Stumpf et al. [116] designed experiments to investigate meaningful explanations and interactions to hold users accountable by machine learning algorithms. They show explanations as a potential for rich human-computer collaboration to share intelligence and keep users accountable.

The recent advancements and trends for explainable AI research demands a wide range of goals for transparency and study across varied application areas. Our review encourages a cross-disicpline perspective of intelligibility and transparency goals.

SURVEY METHOD AND TERMINOLOGY

To help capture and organize the breadth of designs and goals for XAI evaluation, we conducted a survey of the research literature. We used a structured and iterative methodology to find all XAI related scopes and categorize the evaluation methods presented in research articles. Our literature review includes more than 50 papers directly related to the evaluation of XAI systems including but not limited to research on interpretable machine learning, machine explanations in intelligent agents and context-aware systems, explanatory debugging, explainable artificial intelligence, algorithmic transparency and fairness, interactive model visualization, and deep learning visualization. We selected existing works 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 did not want to exclude *arXiv* preprints and useful discussions in workshop papers. In our iterative paper selection method, we started with 40 papers related to

XAI across three research fields as mentioned above, and we used selective coding to identify 15 research attributes. After careful review and analysis of XAI goals and their evaluation methods in the literature, we recognized the following three attributes to be most significant for our purposes of organizing XAI evaluation methods:

- **Research discipline**. We categorized research related to XAI topic into four disciplines including machine learning, data visualization, HCI, and psychology.

- **Targeted Users**. We categorized types of users of XAI systems into three groups: AI novices (most of the general public), data experts (experts in data analytics and domain experts), and AI experts.

- **Evaluation Measures**. The measures used to evaluate the XAI goals. The measures include user's mental model, user's trust and reliance, user's satisfaction and understanding, human-machine task performance, and computational measures.

On the second round of collecting XAI literature, we followed upward and downward literature investigation using the *Google Scholar* search engine to include 30 more papers to our reference table. We focused 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 AI.

With this information, we made a third iteration and used axial coding to organize the literature, and we started discussions on our proposed evaluation categorization. Finally, to maintain reasonable literature coverage and balance the number of papers for each of our categories of evaluation measures and goals, we added another 30 papers to our reference table.

To familiarize readers with common XAI concepts that are repeatedly referenced in this review, the following subsections summarize high-level characterizations of XAI explanations.

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 explanations* (or *model explanations*) in an explanation type that describes how the whole machine learning model works. Model visualization and summaries of decision rules are examples of explanations falling in this category. In contrast, *local explanations* (or *instance explanations*) aim to explain the relationship between a specific input-output pairs or the reasoning behind an individual user queries. 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.

Explanation Formats
As with 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 some researchers have studies forms of explanations that are not limited to humans understanding (e.g., [91]).

Explanations can be designed using a variety of formats. *Visual explanations* use visual elements to describe the reasoning behind the machine learning models. Attention maps and visual saliency in the form of heatmaps [131, 113] are examples of visual explanations that are widely used machine learning literature. Explanation interfaces commonly make use of visual elements combined with various other explanation components. *Verbal explanations* describe the machine learning model and reasoning with words, text, or natural language. Verbal explanations are popular in applications like question-answering explanations, decision lists [67], and explanation interfaces [88]. This form of explanation has also been implemented in recommendation systems[8, 44] and robotics [104] *Analytic explanation* is another approach to see and explore data and machine learning models representations [49] Analytic explanations commonly rely on numerical metrics and data visualizations. For instance, visual analytics methods allow researchers review model structures and their parameters in complex deep models. Heatmap visualizations [115], graphs and networks [36], and hierarchical (decision trees) visualization are commonly used to visualize analytic explanations for interpretable machine learning algorithms.

What to Explain
When users are face a complex intelligent system, they may demand different types of explanatory information. For example, while an explicit model representation of an algorithm is one way to assist in explaining a model, users may also be able to develop a mental model of the system based on a collection of individual instances. Each explanation can be designed to explain in different ways. For instance, *why*-type explanations to describe *why* a result was generated for particular input. Such explanations aim to communicate what features in input data [100, 71] or what logic in the model [101, 67] resulted in a given machine output. Simiarly, *why-not* explanations help the user to understand the reasons why a specific output was not in the machine output [124].

In addition, systems can rely on new queries, changes to the input, or hypothesized inputs to provide *what-if* explanations. *How-to* explanations use the opposite procedure: systems can allow the selection or adjustment of possible outputs and explain hypothetical input conditions to produce that output. Such methods can work interactively with the user's curiosity and partial conception of the system to allow an evolving mental model of system through iterative testing.

Related Surveys
In recent years, there have been surveys and position papers suggesting research directions and highlighting challenges in interpretable machine learning research [75, 45, 40, 72, 29]. Here, we summarize several of the most relevant peer-reviewed surveys related to the topic of XAI across active disciplines. While all surveys in this section add value to the XAI research, to the best of our knowledge, there is no
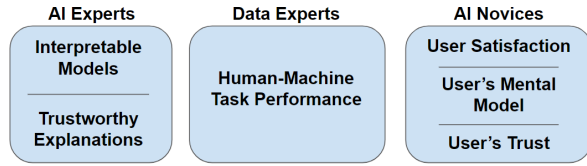
Figure 1. A summary of XAI user types and the most common measures prioritized per user group.

comprehensive survey of evaluation methods for explainable machine learning systems.

Many HCI surveys and reports discuss the limitations and challenges in AI transparency [125, 70] and accountable algorithmic decision-making process [69]. Others suggest a set of theoretical and design principles to support intelligibility of the system and accountability of human users (e.g., [51, 7]). In a recent survey, Abdul et al. [1] presented a thorough literature analysis to find XAI related topics and the relationship between topics. They used visualization of keywords topic model and citation network to present a holistic view of research efforts on XAI from privacy and fairness to intelligent agents and context-aware systems. Further, visualization surveys follow visual analytics goals like understanding and interacting with machine learning systems in different applications [75, 108, 3, 31]. In a recent example, Homan et al. [49] 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. In machine learning, Guidotti et al. [40] present a comprehensive review of machine learning interpretability methods. Also, Montavon et al. [85] and Zhang et al. [132] focus on interpretability in deep neural network models.

Different from all the related surveys mentioned above, our survey provides a multidisciplinary and comprehensive categorization of evaluation methods for XAI systems and presents a model for end-to-end design and evaluation of XAI systems.

CATEGORIZATION OF XAI EVALUATION METHODS

While an ideal XAI system should be able to answer all user queries and meet all XAI concept goals [41], researchers design and study XAI systems with specific explanation goals and specific users. Evaluating the explanations demonstrates or assesses the effectiveness of explanation systems for these goals. In this regard, researchers evaluate the XAI outcomes by different measures such as human-machine task performance, user reliance and trust in machine learning, and accuracy of the user's mental model.

In this section, we categorize XAI evaluation methods based on their targeted users and evaluation measures. We justify our reasoning by describing why and how these two factors affect the evaluation method for XAI systems. However, we note that although our organization is based upon our review of available literature, we do not claim our characterization as perfect or the only way to organize evaluation methodologies. To help summarize the characterization along with example literature, we present a cross-reference table (Table 1) of XAI evaluation literature to emphasize the importance of evaluation

measures and users. It is important to mention that this table does not contain all papers reviewed in this survey.

Targeted Users

Most XAI systems are designed exclusively for a specific explanation type, application, and user. Different explanation characteristics such as explanations type, length, and level of detail will be affected by the explanations purpose and user. For example, while machine learning experts might prefer highly detailed visualization of deep models to help them optimize and diagnose their systems, end users of daily-use AI products do not expect full detailed explanations for every query from a personalized agent. Therefore, we distinguish XAI evaluation subjects into three general groups of machine learning experts, data experts, and AI novices. The reason for this categorization is that XAI systems are expected to provide the right explanations for the right users. Meaning that it is the most efficient to design and evaluate an XAI system according to the user's needs and levels of expertise.

T1: 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 in a more extensive application. XAI systems are expected to respond directly to their end-users with a human-understandable explanation or clarification. In this regard, creating abstract and yet accurate representation of complicated machine learning explanations for novice end-users is a challenge for XAI explanation design.

HCI researchers evaluate XAI systems with multiple measures to find out how to improve the end-user experience. For example, studying users' mental models of the AI is one primary measure for evaluating XAI systems with novices (e.g., [64, 70]). The effects of explanations on end-users' reliance on the system is another common research goal (e.g., [50, 35]). Related to this, an intuitive and user-friendly interactions design can enhance users experience and therefore end-users comprehension and reliance on an XAI system [86].

Both between-subject and within-subject experimental designs are common in evaluating XAI systems. To compare the evaluation outcomes with a baseline, researchers often compare XAI outcomes with AI (i.e., no machine explanations) outcomes. Other times, the comparison is between different explanations types, complexity, and formats. There are various metrics commonly used in research to measure users' mental model and trust through human subjects studies. User performance on human-machine tasks, user precision and recall in predicting machine outputs, user trust and confidence on machine predictions, and user success in predicting AI outcomes are examples of measures used in evaluation. Below, we briefly introduce two evaluation types commonly used in HCI research to evaluate machine learning explanations.

Table 1. Tabular summary of our XAI evaluation dimensions of measures and targeted user types. The table includes a representative subset of the surveyed literature organized by the two dimensions. Although, most papers study machine explanations with more than one measure, to reduce clutter and redundant references numbers, we listed references based on their primary evaluation measure.

Interpretability Measures	Data Novices	Data Experts	Machine Learning Experts
User's Mental Model	[64] [67] [70] [100] [102] [117] [44]	–	–
Human-Machine Task Performance	[39] [63] [65] [100] [88] [124] [116] [71]	[62] [25] [19][36] [53] [73] [20] [14] [76] [107] [9] [61]	[115] [83] [75] [94] [55] [133] [74]
User Satisfaction of Explanation	[10] [71] [34] [89] [15] [66] [112]	–	–
User Trust and Reliance	[50] [8] [97] [90] [35] [43] [5] [50]	[16]	–
Computational Measures	–	–	[131] [59] [105] [100] [22] [110] [101] [129]

In-lab studies

Lab-based studies involve of inviting human participants to collect feedback regarding users experience with the XAI system. User feedback can be collected in various forms such as system logs, questionnaires, self-reports, and think-aloud reports. Although users of in-lab studies provide high-quality data and feedback about their experience with the XAI system, the downside of in-lab studies is that running this type of study is time-consuming and difficult to scale to a large number of participants.

Online and crowdsourced studies

Online studies provide the benefit of collecting data from study participants without requiring schedule coordination between an experimenter and participant, and participants are often able to complete the study at their convience of time and location. Crowdsourcing is the key to scalable online data collection from human-users [60]. In a crowdsourced XAI evaluation design, remote users would participate in the study to review machine explanations and provide feedback. Many data collection methods are still possible (e.g., performance logs, questionnaires, and self-reports) in crowdsourced studies, but methods such as interviews or not often feasible. It is also advisable that online studies are designed to hold the attention of participants for the duration of the study, as experimenter monitoring and intervention are not possible. As a result, online studies often use micro-tasks to increase data quality. Other methods to ensure data quality in crowdsourcing studies are manipulation checks and attention checks during the study as a form of data validation.

T2: Data Experts

Data experts include data scientists and domain experts who use machine learning for analysis, decision-making, or research. This group of targeted users also includes researchers or workers who analyze data in specialized domains such as cybersecurity [36, 9], medical [19, 62], text analysis [76, 73],

and image analysis [103]. These users might be experts of certain domain areas or experts in general areas of data science, but for of our organization, we consider users in the *data experts* who generally lack expertise in the technical specifics of AI or machine learning algorithms. This group of users often uses interactive data analysis tools, recommender systems, or visual analytics systems that combine interactive interfaces and algorithms.

Data experts also benefit from machine explanations to inspect uncertainty and investigate algorithms prediction accountability. For example, machine-learning explanations help data experts to find problems with training-bias in supervised machine learning models. Therefore, a main challenge for data-analysis and decision-support systems is to increase model transparency and user awareness with visualization and interaction techniques [109]. Visual analytics approaches can help data experts to tune machine learning parameters for their specific data in an interactive visual fashion. Visualizing details and explanations of machine learning output may result in better understanding the machine algorithms behavior [76].

Similar to evaluations with AI novices, evaluating analytics tools for data knowledgable users and domain experts often involves human subjects. However, many intepretable analytics tools are designed for data and machine learning experts. Visual analytics expert evaluations enter when controlled experiments fail due to high cognitive tasks [122]. In practice, it can be difficult to gain access or take the time of large numbers of experts for evaluation, which often makes it difficult to evaluate with controlled studies.

Case studies, expert reviews, and focus groups are often used for evaluation of visual analytics and decision-support tools for experts [95, 18]. Case studies aim to collect experts users feedback while performing high-level cognitive tasks. Expert review and interview sessions in case studies involve informal question answering and experts think-aloud to measure experts satisfaction. After interviewing experts individually or in

the form of focus groups, analyzing the experts' thoughts, agreements, and disagreements is useful for informing design and assessing usability issues. Experts can tell us how well the current design can help them in understanding complex machine learning models, and these reviews are also valuable parts of the visual analytics design cycle [87, 121].

T3: Machine Learning Experts

Machine learning experts are scientists and engineers who design interpretable machine learning algorithms as well as other machine learning algorithms. Various visualizations and visual analytics tools help these machine learning experts to verify the model accuracy [74, 55]. However, machine-generated explanations can also be evaluated with computational methods (rather than human-subject review) especially to validate the explanation trustworthiness. Computational evaluations are common in the field of machine learning and focus on measuring the accuracy of the explanations in terms of mirroring what the model has learned. Various computational experiments have been used to measure machine-generated explanations faithfulness to the model, as well as evaluating explanation reliability by simulating a real user.

The main goal of computational evaluations is to measure explanations truthfulness in terms of describing model behavior, mirroring what the model has learned, and assessing general computational performance. The main challenge is that there is no ground-truth explanation available for black-box models to compare the results.

To name primary measures used in literature for computational evaluation of explainable machine learning we can mention, the accuracy of explainable models, generating explanation's speed, explanations consistency, explanations precision and recall (comparing to an interpretable model), and comparison with state-of-the-art explanation methods. There are also examples of simulating human users to evaluate users trust in machine-generated explanations.

Data analysis tools also can support interpretable machine learning in many ways such as visualizing network architecture and inspecting learned model parameters [49]. Researchers have implemented various visualization and interaction designs to better understand and improve machine learning models. Visual analytics designs have also been developed to help machine-learning novices to learn deep models by interacting with simplified representations of models [114].

Evaluation Measures

In addition to user type, *evaluation measure* is another important factor in the design and evaluation of XAI systems. Explanations are designed to answer different interpretability goals, and different measures can verify explanations goodness for that purpose. Since a system's evaluation plan should match the design goals, researchers evaluate XAI goals with different measures and metrics. For instance, study designs to evaluate users trust in XAI systems take different metrics from evaluation plans for measuring explanations quality. Table 1 shows a list of five evaluation measure that we extracted from the literature. We categorize the main XAI evaluation measures in the following sections.

M1: Mental Model

Following cognitive psychology, a mental model is a representation of how a user understands a system. Researchers study users' mental models to determine user understanding of intelligent systems in various applications. For example, Costanza et al. [23] studied how users understand a smart grid system, and Kay et al. [56] studied how users understand and adapt to uncertainty in machine learning prediction of a bus arrival times. In the context of XAI, explanations help users to create a mental model of *how the AI works*. Machine explanation is a method to help the user in building a more accurate mental model. Psychology research has also explored structure, types, and functions of explanations to find essential ingredients of ideal explanation for better user understanding and mental models [77, 57, 78]. In order to find out how an intelligent system should explain its behavior for non-experts, research on machine learning explanations studied how users interpret intelligent agents [28, 93] and algorithms [99] to find out what users expect from machine explanations [70, 28].

A useful way of studying user comprehension of intelligent systems is to directly ask the user about the intelligent system's decision-making process. Analyzing user interviews, think alouds, and self-explanations provides valuable information about user thought process and mental model [65]. On studies of how explanations complexity affect user comprehension, Kulesza et al. [66] studied the impact of explanations soundness and completeness on the fidelity of the end users mental model in a music recommendation interface. Their results found explanation completeness (broadness) had a more significant effect on user understanding of the agent compared to explanation soundness. User attention and expectations when using intelligent systems may also be considered as means used as implicit methods for assessing user understanding and approximating a user's mental model [117]. In another example, Binns et al. [11] studied the relation between machine explanations and user perception of justice in algorithmic decision-making with different sets of explanations style.

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. [100] evaluated users understanding of machine learning algorithm with visual explanations. They showed how explanations mitigate users overestimation of an image classifier and choose the better classifier based on their explanations. In followup 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 [102]. In another paper, Kim et al. [58] conducted a crowdsourced study to evaluate feature-based explanation understandability for end-users. In the understanding of a machine learning model representations, Lakkaraju et al. [67] presented interpretable decision sets, an interpretable classification model, and measured users' mental models with different metrics such as users accuracy on predicting machine output and length of users' self-explanations.

M2: Human-Machine Task Performance

A key goal of XAI is to help end-users to be more successful in their tasks involving machine learning systems [51]. Thus, human-machine task performance is a measure relevant to all three groups of user types. For example, Lim et al. [71] measured user performance in terms of completion time and test accuracy to evaluate the impact of different types of explanations. They showed machine explanations to have a significant impact on users' accuracy in their task. Also, explanations can assist users in adjusting the intelligent system to their needs. Kulesza et al. [64] 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 and experience with the system.

Another use case for machine learning explanations is to help users to justify the correctness of system output [39, 61, 116]. Explanations also assist users in debugging interactive machine learning programs for their needs [63, 65]. In a study of end-users interacting with a email classifier system, Kulesza et al. [63] measured users' mental models and classifiers performance to show explanatory debugging benefits both user and machine performance. Similarly, Ribeiro et al. [100] found users could to detect and remove wrong explanations in text classification, resulting in training better classifiers by rewiring the algorithms and changing its logic. To support these goals, Myers et al. designed a framework that users can ask why and why not questions and expect explanations from intelligent interface [88].

Visual analytics tools also help domain experts to outperform in their tasks by providing model interpretation. Visualizing model structure, details, and uncertainty in machine outputs can allow domain experts to diagnose models and adjust parameters to their specific data for better analysis. Visual analytics research explored the need for model interpretation in text [126, 53, 73] and multimedia [20, 14] analysis tasks, and they demonstrate the importance of integrating user feedback to improve results. An example of a visual analytics tool for text analysis is TopicPanaroma [76] that models a textual corpus as a topic graph and incorporates metric 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 relations manager used the tool to find a set of tech-related patterns in news media and a professor analyzing the news media impacts on public during a health crisis.

In analysis of streaming data also automated approaches in are error-prone and require expert users to review model details and uncertainty for better decision making [107, 9]. Goodall et al. [36] presented Situ, a visual analytics system for discovering suspicious behavior in 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.

Other than domain experts using from visual analytics tools, machine learning experts also use visual analytics to find model architecture shortcomings or training flaws in deep neural networks to improve classification and prediction performance [75, 94]. For instance, the LSTMVis [115] and RNNVis [83] are tools to interpret RNN models for natural language processing tasks. In a recent example, Kahng et al. [55] designed a VA system to visualize instance-level and subset-level of neuron activations in a long term investigation and development with machine learning engineers. In their case studies, they interview three Facebook engineers and data scientist to use the tool and reported the key observations. Another critical role of visual analytics for machine learning experts is to visualize model training processes [133, 74]. An example of a visual analytics tool for diagnosing the training process of a deep generative model is DGMTracker by Liu et al. [74]. DGMTracker helps experts understand the training process by represents training dynamics. They conducted two case studies with experts to validate the efficiency of DGM-Tracker in understanding the training process and diagnosing a failed training process.

M3: Explanation Satisfaction

End-user satisfaction and understanding of machine explanation is another measure to evaluate explanations in intelligent systems [10]. Researchers use different subjective factors such as understandability, usefulness, sufficiency of details, and accuracy to measure the explanatory value of the machine explanations to users [82]. Although there are implicit methods to measure user satisfaction [46], most literature follows qualitative evaluations of explanations satisfaction with questionnaire and interview methods. For example, Gedikli et al. [34] evaluated ten different explanation types with users rating of explanations satisfaction and transparency. Their results showed a strong relationship between user satisfaction and perceived transparency. Similarly, Lim et al. [71] evaluated explanation understandability and learning time on the system by presenting different types of explanations like "why", "why not" and "what if" explanations. To study the impact of explanations complexity on users comprehension, Narayanan et al. [89] studied how explanation length and complexity affect user response time, accuracy, and subjective satisfaction. To investigate explanations in real-world use case, Lim and Dey [70] also studied user understanding and satisfaction of different explanation types in four real-world context-aware applications. In contrast, Bunt et al. [15] considered whether explanations are always necessary for users in every intelligent system. Their results show that the cost of viewing explanations in diary entries like Amazon and YouTube recommendations could be more than their benefit.

Note that, for simplicity, our organization does not explicitly include direct user satisfaction and verification of explanations as primary goals of evaluations of interpretable systems for data experts, user comprehension of visualizations and interactions usability are always considered as a part of the evaluation pipeline during interviews and case studies.

M4: User Trust and Reliance

User trust in an intelligent system is an affective and cognitive element that influences positive or negative perceptions of a system [79, 48]. Initial user trust and the development of trust over time have been studied and presented with different terms

such as *swift* trust [81], *default* trust [80] and *suspicious* trust [12]. Prior knowledge and beliefs are important in shaping the initial state of trust; however, trust and confidence can update in response to exploring and challenging the system with edge cases [47]. 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 [79, 54]. 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. Additionally, trust assessment scales could be specific to the systems application context and XAI design purposes. For example, multiple scales would assess user opinion on systems reliability, predictability, and safety separately. An example of a detailed trust measurement setup is presentation in work by Cahour and Forzy [17], which measures users 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 user trust in adaptive agents, Glass et al. [35] 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 [5, 56] .

Multiple efforts have studied the impact of XAI on developing justified trust in users in different domains. For instance, Pu and Chen [97] proposed an organizational framework for generating explanations. They measured perceived competence and intention to return as measures for user trust. Another example compared user trust with explanations for different goals like transparency and justification explanation [90]. 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.

Evaluating user trust in real-world applications, Berkovsky et al. [8] evaluated trust with various recommendation interfaces and content selection strategies. They evaluated user reliance on a movie recommender system with six distinct constructs of trust. Also concerned with expert trust, Bussone et al. [16] measured trust by Likert-scale and think-alouds. They found explanations of facts lead to higher user trust and reliance in a clinical decision-support system.

Many studies evaluate the user trust as a static state. However, it is essential to take users experience and learning over time into account to measure the user's trust. Collecting repeated measures over time can help in understanding and analyze the trend of users developing trust with the progression of experience. For instance, in their trust assessment study, Holliday et al. [50] evaluated trust and reliance in multiple stages of working with an explainable text-mining system. They showed 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 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 evaluation pipeline during interviews and case studies. In other words, our summarization is not meant to claim that data experts do not consider trust, but rather that we did not find it to be a core evaluation goal in the literature for this user group.

M5: Computational Measures

Machine-generated explanations can also be evaluated by computational methods instead through human-subject studies. Computational methods are prevalent in AI and machine learning research to verify explanation correctness and usefulness. However, in many cases, machine learning researchers often consider model consistency, computational interpretability, and self-interpretation of their results as evidence for explanation correctness [91, 129, 134]. In other cases, Zeiler and Fergus [131] discuss how the visualization of convolution neural networks can help find model weaknesses and obtain better prediction results. Another example is the Deep Visualization Toolbox [128], an interactive tool to explore convolutional neural networks. This toolbox visualizes activation layers of a convolution neural network in real-time and gives intuition about "how the system works" to the user.

Others study trustworthiness for edge cases where these methods give unreliable explanations [59, 105]. An example of such investigations is Kindermans et al. [59] work on unreliability of explanations from saliency methods. Their findings show that to achieve reliable explanations from saliency methods, the methods should fulfill input invariance to mirror the sensitivity of the model itself. In some cases, comparing a new explanation method with other state-of-the-art explanation methods (e.g., LIME [100]) is a way to verify explanation quality [22]. In other work, Ross et al. [106] designed a comprehensive set of empirical evaluations to compare their explanations' consistency, features, and generation speed with the LIME ethod [100]. Several methods including simulation experiments and explanations precision comparison have been used to measure machine generated explanations stability and accuracy. For example, Ribeiro et al. [100] compared explanations generated by their ad-hoc explainer to explanations from an interpretable model. They created gold standard explanations directly from the interpretable (sparse logistic regression and decision trees) models and used these for comparisons in their study. The downside of this that evaluation is limited to generating gold standard by an interpretable model.

In a different approach, Samek et al. [110] proposed a framework for evaluating heatmap explanations for image data that quantify the importance of pixels with respect to the classifier prediction. They compared three different heat-map explanation methods for image data (sensitivity-based [113], deconvolution [131], and layer-wise relevance propagation [6]) and investigate the correlation between heatmap quality and network performance on different image datasets.

User-simulated evaluation is another method to perform computational evaluations of machine-generated explanations.

Ribeiro et al. [100] simulate users trust on the explanations and models by defining "untrustworthy" explanations and models. They tested a hypothesis about how real users would prefer more reliable explanations and choose better models. The authors later repeated similar user-simulation evaluations in their Anchors explanation approach [101] to report simulated users precision and coverage in finding better classifier by only looking at explanations.

DISCUSSION

In our review, we discussed multiple XAI evaluation measures appropriate for various targeted user types. Table 1 presents a 2D categorization of existing evaluation methods that organizes literature into along two perspectives: *evaluation measure* and *targeted users* of the XAI system. Now, we summarize considerations for a comprehensive evaluation of end-to-end explainable intelligent systems followed by a nested evaluation model for XAI systems (Figure 2).

Given that machine learning, visual analytics, and human-computer interaction research disciplines actively work in design and evaluation of interpretable intelligent systems, XAI evaluation should be interdisciplinary effort that takes consideration for both human and computational elements.

Considerations for XAI Evaluations

In this section, we discuss considerations for XAI designers to benefit from the body of knowledge on XAI design and evaluation. Similar to user-centered design principles [2], we suggest considering design and evaluation of explainable intelligent systems as the usability engineering cycle to increase expected outcomes.

As research communities have different priorities in design and evaluation of interpretable machine learning systems, we suggest to begin the XAI evaluation plan with *identifying the intended user*. As discussed in the prior sections, studies involving XAI applications for novice end-users have a separate path than applications for domain experts and also from computational measures of interpretability.

Next is to *choose an XAI application* that consists of machine learning explanations to benefit the targeted user's primary goals and needs. This may also affect the underlying interpretable machine learning model, interpretable interface design, explanation format, and even measure details like choosing right constructs of trust. However, in some cases, evaluation plans limited to studying specific design factors and does not necessarily require a practical application. It is crucial to choose the application domain at the early stages of the study [87].

Depending on the XAI application, user type, and underlying machine learning models, it is then possible to *choose the explanation type and format* to continue with the design plan. Explanation format also usually has restrictions for the interpretable model's or ad-hoc explainer's output format. Similar to user-centered design, at this stage, one can perform preliminary evaluations of user satisfaction and understandability of explanations. Possible factors to study are including explanations format (e.g., verbal, visual, or analytical), type (e.g.,

why, why not, what, how), and the appropriate level of detail. At this step, explanation interfaces and agents translate machine-generated explanations to a user-understandable (and satisfactory) content for the end-user.

Based on goals and expected outcomes of an XAI system, one would choose higher-level measures to study explanations. In some situations, however, the study may prioritize achieving sufficient satisfaction ratings above all else, so it may not require measuring deeper forms of user understanding and feelings about XAI. As found in our review, user trust and mental models are common measures for applications targeting novice users. These applications include explainable recommendations systems, personal agents and intelligent interfaces. However, human-machine task performance is commonly used in applications for domain and AI experts. Visual analytics tools are often designed for users with specific types expertise such as data scientists, deep learning engineers, and other domain experts (e.g., medical physicians, cybersecurity experts, economic analysts). In other circumstances, a novice user's mental model may result in better operation of intelligent systems, thus leading to outcomes. For example, explanations in recommendation systems may not only improve user's mental model and trust in the system, but also lead to improved overall decision making and higher user satisfaction.
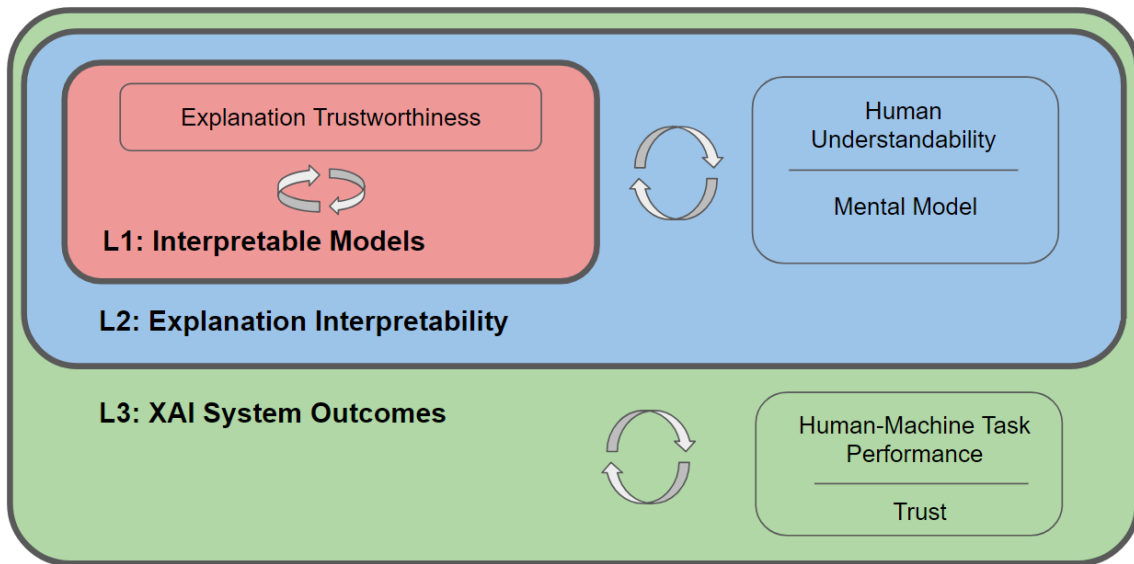
XAI Design and Evaluation Model

Following our review and organization, we now present a multilayer model for XAI system design and evaluation. The impetus for this evaluation model was the desire to organize and relate the diverse set of existing XAI design and evaluation papers that cover a broad set of goals. With this model, we aim to further facilitate an interdisciplinary path to XAI research and system design. The model is intended to give guidance on what evaluation measures are appropriate to use at which stage of an XAI design. Taking an emphasize on an efficient design and evaluation pipeline, we split the XAI evaluation process into three levels:

1. Interpretable Models
2. Explanations Interpretability
3. XAI System Outcomes

Figure 2 summarizes our three-level model for end-to-end XAI system design and evaluation.

The innermost layer
Understandable models are the core foundation of an XAI system. Interpretability is what we aim to deliver to the end-users, and we need to achieve it in the back-end of the system. The goal for the innermost layer is to design interpretable models (or generate machine explanations) and verify explanations trustworthiness. Machine learning experts contribute by designing interpretable models (or ad-hoc explainers) with distinct features and explanation types. Machine learning experts also evaluate the explanations' trustworthiness (or fidelity of ad-hoc explainer to the black-box model) with computational measures. The interpretable machine learning designers are not only engaged in the interpretability and accuracy trade-off but also are involved with explanations trustworthiness and

Figure 2. XAI Evaluation Model: our nested evaluation model for evaluating explainable machine learning systems. The innermost layer (Red) presents design and evaluation of interpretable machine learning algorithms. The middle layer (Blue) shows design and evaluates human understandable explanations and explainable intelligent interfaces and agent. The outer layer (Green) demonstrates evaluation of XAI system outcomes with end-users.

reliability. This layer is essential because any unreliability of interpretability at this inner layer will propagate to all other outer layers.

The middle layer
This layer can be viewed as the translator for an XAI system. An elegant translation (i.e., verbal, visual, or visualization) of machine-generated explanations is what we aim to deliver to the end-users, and we need to present it at the front-end of the system. The goal of the middle layer is to design human understandable explanations, explanation agents and interfaces, and visual analytics tools and verify their usability. HCI and visual analytics designers employ user-centered design principles and verify designs with subjective evaluations of satisfaction and mental models for different types of users.

The outer layer
The outermost layer focuses on *outcomes* of using the XAI system. The aim of this layer is to verify what the intelligible design has gained for the end-user. In other words, the goal is to measure XAI system's impact on user trust, reliance, and human-machine performance in performing tasks. Evaluation designs in this layer are very much dependent on the application domain, and various subjective and objective measures can be used to evaluate XAI system success with regard to the expected outcomes.

CONCLUSION
We reviewed XAI-related research to organize evaluation methods and show a mapping between user groups and evaluation measures, and we proposed a nested model for the design and evaluation of XAI systems. The model can serve as a reference when designing and evaluation of new interpretable systems or for assessing the strengths and limitations of evaluations. For example, evaluating a newborn interpretable machine learning algorithm's output using human subjects through a weak

user or crowdsourcing study may not be meaningful or productive if core computational changes are still in progress that would ultimately change the entire model interpretability and explanation format later. It may also not be prudent to expect improvements to interpretable algorithms to directly improve user's trust or task performance without revisiting the explanations and the impact on end-user outcomes.

The model highlights that interpretability issues in the inner layers will inevitably propagate to the outer layers. In the end, we want to further emphasize the need to consider machine explanation as a process over time for human-users rather than a constant factor. The user's process of shaping and refining the mental model includes a series of interactions with the explainer followed by learning system and unlearning misconceptions. We hope this model drives further discussion about the interplay between design and evaluation of explainable artificial intelligent systems.

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. Chadia Abras, Diane Maloney-Krichmar, and Jenny Preece. 2004. User-centered design. *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications* 37, 4 (2004), 445–456.

3. 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.

4. 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.

5. 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.

6. 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.

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

8. 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. DOI: http://dx.doi.org/10.1145/3025171.3025209

9. 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.

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

11. 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.

12. 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.

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

14. 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.

15. 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.

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

17. 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.

18. Sheelagh Carpendale. 2008. Evaluating information visualizations. In *Information visualization*. Springer, 19–45.

19. 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.

20. 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.

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

22. 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. *arXiv preprint arXiv:1802.06259* (2018).

23. 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.

24. 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.

25. Anind K. Dey and Alan Newberger. 2009. Support for Context-aware Intelligibility and Control. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 859–868. DOI: http://dx.doi.org/10.1145/1518701.1518832

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

27. 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.

28. Jonathan Dodge, Sean Penney, Andrew Anderson, and Margaret Burnett. 2018. What Should Be in an XAI Explanation? What IFT Reveals. (2018).

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

30. Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. 2017. Accountability of AI Under the Law: The Role of Explanation. *arXiv preprint arXiv:1711.01134* (2017).

31. 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.

32. 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.

33. 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 *ICWSM*. 62–71.

34. 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.

35. 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.

36. 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).

37. Bryce Goodman and Seth Flaxman. 2016. EU regulations on algorithmic decision-making and a âĂIJright to explanationâĂİ. arXiv preprint. *arXiv preprint arXiv:1606.08813* (2016).

38. Shirley Gregor and Izak Benbasat. 1999. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly* (1999), 497–530.

39. Alex Groce, Todd Kulesza, Chaoqiang Zhang, Shalini Shamasunder, Margaret Burnett, Weng-Keen Wong, Simone Stumpf, Shubhomoy Das, Amber Shinsel, Forrest Bice, and others. 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.

40. 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.

41. David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* (2017).

42. 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.

43. 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.

44. 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.

45. Bernease Herman. 2017. The Promise and Peril of Human Evaluation for Model Interpretability. *arXiv preprint arXiv:1711.07414* (2017).

46. Robert R Hoffman. 2017. TheoryâĘŠ conceptsâĘŠ measures but policiesâĘŠ metrics. In *Macrocognition Metrics and Scenarios*. CRC Press, 35–42.

47. 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.

48. Robert R Hoffman, Matthew Johnson, Jeffrey M Bradshaw, and Al Underbrink. 2013. Trust in automation. *IEEE Intelligent Systems* 28, 1 (2013), 84–88.

49. 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).

50. 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.

51. Kristina Höök. 2000. Steps to take before intelligent user interfaces become real. *Interacting with computers* 12, 4 (2000), 409–426.

52. Philip N Howard and Bence Kollanyi. 2016. Bots,# StrongerIn, and# Brexit: computational propaganda during the UK-EU referendum. (2016).

53. Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2014. Interactive topic modeling. *Machine learning* 95, 3 (2014), 423–469.

54. 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.

55. Minsuk Kahng, Pierre Y Andrews, Aditya Kalro, and Duen Horng Polo Chau. 2018. A cti V is: Visual Exploration of Industry-Scale Deep Neural Network Models. *IEEE transactions on visualization and computer graphics* 24, 1 (2018), 88–97.

56. 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.

57. Frank C Keil. 2006. Explanation and understanding. *Annu. Rev. Psychol.* 57 (2006), 227–254.

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

59. Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2017. The (Un) reliability of saliency methods. *arXiv preprint arXiv:1711.00867* (2017).

60. Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 453–456.

61. Josua Krause, Aritra Dasgupta, Jordan Swartz, Yindalon Aphinyanaphongs, and Enrico Bertini. 2017. A Workflow for Visual Diagnostics of Binary Classifiers using Instance-Level Explanations. *arXiv preprint arXiv:1705.01968* (2017).

62. 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.

63. 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.

64. 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. DOI: `http://dx.doi.org/10.1145/2207676.2207678`

65. 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.

66. 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.

67. 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.

68. 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.

69. 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.

70. 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.

71. 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.

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

73. 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.

74. 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.

75. 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.

76. 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.

77. Tania Lombrozo. 2006. The structure and function of explanations. *Trends in cognitive sciences* 10, 10 (2006), 464–470.

78. Tania Lombrozo. 2009. Explanation and categorization: How âĂIJwhy?âĂİ informs âĂIJwhat?âĂİ. *Cognition* 110, 2 (2009), 248–253.

79. Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *11th australasian conference on information systems*, Vol. 53. Citeseer, 6–8.

80. Stephanie M Merritt, Heather Heimbaugh, Jennifer LaChapell, and Deborah Lee. 2013. I trust it, but I donâĂŹt know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human factors* 55, 3 (2013), 520–534.

81. 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.

82. Tim Miller. 2017. Explanation in artificial intelligence: insights from the social sciences. *arXiv preprint arXiv:1706.07269* (2017).

83. Yao Ming, Shaozu Cao, Ruixiang Zhang, Zhen Li, Yuanzhe Chen, Yangqiu Song, and Huamin Qu. 2017. Understanding Hidden Memories of Recurrent Neural Networks. *arXiv preprint arXiv:1710.10777* (2017).

84. Brent Mittelstadt. 2016. Automation, Algorithms, and Politics| Auditing for Transparency in Content Personalization Systems. *International Journal of Communication* 10 (2016), 12.

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

86. 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.

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

88. 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.

89. Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *arXiv preprint arXiv:1802.00682* (2018).

90. 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.

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

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

93. 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. `DOI:` `http://dx.doi.org/10.1145/3172944.3172946`

94. 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.

95. Catherine Plaisant. 2004. The challenge of information visualization evaluation. In *Proceedings of the working conference on Advanced visual interfaces*. ACM, 109–116.

96. 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.

97. 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.

98. 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.

99. 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.

100. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 1135–1144.

101. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018a. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence*.

102. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018b. Anchors: High-Precision Model-Agnostic Explanations. In *AAAI Conference on Artificial Intelligence (AAAI)*.

103. 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.

104. Stephanie Rosenthal, Sai P Selvaraj, and Manuela M Veloso. 2016. Verbalization: Narration of Autonomous Robot Experience.. In *IJCAI*. 862–868.

105. Andrew Slavin Ross and Finale Doshi-Velez. 2017. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. *arXiv preprint arXiv:1711.09404* (2017).

106. 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. DOI: **http://dx.doi.org/10.24963/ijcai.2017/371**

107. Stephen Rudolph, Anya Savikhin, and David S Ebert. 2009. Finvis: Applied visual analytics for personal financial planning. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*. Citeseer, 195–202.

108. Dominik Sacha, Michael Sedlmair, Leishi Zhang, John Aldo Lee, Daniel Weiskopf, Stephen North, and Daniel Keim. 2016a. Human-centered machine learning through interactive visualization. ESANN.

109. Dominik Sacha, Hansi Senaratne, Bum Chul Kwon, Geoffrey Ellis, and Daniel A Keim. 2016b. The role of uncertainty, awareness, and trust in visual analytics. *IEEE transactions on visualization and computer graphics* 22, 1 (2016), 240–249.

110. 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.

111. 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.

112. 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.

113. 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).

114. 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).

115. 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.

116. 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.

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

118. Latanya Sweeney. 2013. Discrimination in online ad delivery. *Queue* 11, 3 (2013), 10.

119. 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.

120. Nava Tintarev and Judith Masthoff. 2011. Designing and evaluating explanations for recommender systems. In *Recommender systems handbook*. Springer, 479–510.

121. Melanie Tory and Torsten Moller. 2004. Human factors in visualization research. *IEEE transactions on visualization and computer graphics* 10, 1 (2004), 72–84.

122. Melanie Tory and Torsten Moller. 2005. Evaluating visualizations: do expert reviews work? *IEEE computer graphics and applications* 25, 5 (2005), 8–11.

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

124. 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.

125. Adrian Weller. 2017. Challenges for transparency. *arXiv preprint arXiv:1708.01870* (2017).

126. 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.

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

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

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

130. 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.

131. Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 818–833.

132. 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.

133. 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*.

134. 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).