Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations

Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci

(Survey Paper)

***Abstract*—Explainable AI (XAI) is widely viewed as a sine qua non for ever-expanding AI research. A better understanding of the needs of XAI users, as well as human-centered evaluations of explainable models are both a necessity and a challenge. In this paper, we explore how human-computer interaction (HCI) and AI researchers conduct user studies in XAI applications based on a systematic literature review. After identifying and thoroughly analyzing 97 core papers with human-based XAI evaluations over the past five years, we categorize them along the measured characteristics of explanatory methods, namely *trust, understanding, usability*, and *human-AI collaboration performance*. Our research shows that XAI is spreading more rapidly in certain application domains, such as recommender systems than in others, but that user evaluations are still rather sparse and incorporate hardly any insights from cognitive or social sciences. Based on a comprehensive discussion of best practices, i.e., common models, design choices, and measures in user studies, we propose practical guidelines on designing and conducting user studies for XAI researchers and practitioners. Lastly, this survey also highlights several open research directions, particularly linking psychological science and human-centered XAI.**

***Index Terms*—Explainable AI (XAI), human-centered XAI, explainable ML, user study, human-AI interaction.**

I. Introduction

Artificial Intelligence (AI) is driving digital transformation and is already an integral part of various everyday technologies. Recent developments in AI are essential to progress in fields such as recommendation systems [97], [98], [99], autonomous driving [100], [101], [102] or robotics [103], [104], [105]. Moreover, AI's success story has not excluded high-stakes decision-making tasks like medical diagnosis [106], [107], [108], credit scoring [109], [110], [111], jurisprudence [112], [113] or recruiting and hiring decisions [114], [115], However, the behavior and decision-making processes of modern AI systems are often not understandable, so they are frequently considered black boxes. Deploying such black-box models presents a serious dilemma in certain safety-critical domains, for instance, public health or finance [116]. This is due to the necessity for a transparent and trustworthy AI system, which is required by both practitioners (to gain better insights into system functioning) and end users (to rely on model decisions).

Methods to increase the interpretability and transparency of an AI system are developed in the research area of Explainable AI (XAI). Specifically, human-centered XAI, which addresses the importance of human stack-holders to the AI systems, has been proposed and discussed since [117], [118]. While a huge number of model explanations are available, the question of how to transparently evaluate their quality is still an open research question, and hence, extensively studied in recent years. A popular taxonomy of evaluation strategies for XAI methods proposes three categories: functionally-grounded evaluation, application-grounded evaluation, and human-grounded evaluation [119]. While functionally-grounded measures do not require human labor, the other two involve human subjects and are more costly to conduct.

Many functionally-grounded measures have been proposed to evaluate XAI algorithms (see [120] for review), however, the difficult comparability between different automatic evaluation measures is a common problem [121], [122]. Another drawback of automated measures is that there is no guarantee that they truly reflect humans' preferences [40], [123]. Consequently, user studies in XAI, especially when moving towards real-world products, are inevitable if one wishes to test more general beliefs of the quality of explanations [16]. However, only a small portion (about 20%) of XAI evaluation projects consider human subjects [120]. There exist efforts in developing taxonomies or introducing the definitions or implications of different human-centric evaluations [124], [125], [126], but the recent generation of user studies and their findings have not been systematically discussed yet. Moreover, Yang et al. [127] point out that XAI is growing separately and treated differently in different communities (e.g., machine learning and HCI). Hence, effective guidance in XAI user study design is crucial to better let both XAI algorithm

Manuscript received 3 February 2023; revised 26 October 2023; accepted 4 November 2023. Date of publication 13 November 2023; date of current version 6 March 2024. Recommended for acceptance by M. Cheng. *(Corresponding author: Yao Rong.)*

Yao Rong, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci are with the Technical University of Munich, 80335 Munich, Germany (e-mail: yao.rong@tum.de; tina.seidel@tum.de; gjergji.kasneci@tum.de; enkelejda.kasneci@tum.de).

Tobias Leemann, Thai-Trang Nguyen, and Lisa Fiedler are with the University of Tübingen, 72076 Tübingen, Germany (e-mail: tobias.leemann@uni-tuebingen.de; thai-trang.nguyen@student.uni-tuebingen.de; lisa.fiedler@student.uni-tuebingen.de).

Peizhu Qian and Vaibhav Unhelkar are with the Rice University, Houston, TX 77005 USA (e-mail: pq3@rice.edu; vaibhav.unhelkar@rice.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2023.3331846, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2023.3331846

TABLE I
OVERVIEW OF THE CORE PAPERS CONTAINING USER STUDIES IN XAI GROUPED BY CATEGORIES OF MEASUREMENTS AS SOME CORE PAPERS ASSESS QUANTITIES BELONGING TO SEVERAL GROUPS, A SINGLE PAPER CAN ALSO BE LISTED AMONG MULTIPLE GROUPS

Trust		[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15] [16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31]
Understanding	subjective	[7, 12, 13, 14, 16, 17, 22, 28, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44]
	objective	[12, 13, 22, 32, 35, 39, 40, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60]
	explanation model	[21, 46, 49, 61, 62, 63, 64, 65]
Usability	workload	[3, 16, 21, 48, 66]
	helpfulness	[13, 45, 46, 48, 56, 65, 67, 68]
	satisfaction	[1, 6, 7, 16, 18, 19, 29, 47, 69, 70]
	undesired behavior detection	[2, 24, 27, 38, 53, 57, 71, 72, 73, 74, 75, 76, 77, 78, 79]
	ease of use and others	[1, 3, 13, 20, 21, 24, 30, 32, 37, 48, 65, 66, 71, 80, 81, 82, 83, 84, 85, 86, 87]
Human-AI Collaboration Performance		[10, 13, 15, 25, 25, 29, 30, 39, 43, 53, 56, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97]

and application designers recognize the users' real needs. This work aims to bridge this research gap in modern XAI user study design by distilling practical guidelines for user studies through a comprehensive and structured literature review.

Therefore, we reviewed highly relevant papers that include user studies from top-tier HCI and XAI venues. Specifically, we included the recent *five* years of CHI, IUI, UIST, CSCW, FA(cc)T, ICML, ICRL, NeurIPS, and AAAI. As we aim at analyzing human user evaluation of advanced model explanations, we ran search queries involving keywords from the two groups "explainable AI" and "user study", as listed in the Table II. We selected the papers containing at least one keyword from each group, resulting in over one hundred papers. Then, we thoroughly studied these papers and filtered out papers that did not fulfill the criteria: (1) deploying explainable models or techniques and (2) conducting an assessment with human subjects. We identified a total of 97 core papers for this survey (see Table I for an overview of core papers with respect to their measured quantities in user studies). Based on these core papers, we performed a comprehensive analysis to fill the research gap by offering a systematic overview of user studies in XAI. We highlight the main contributions:

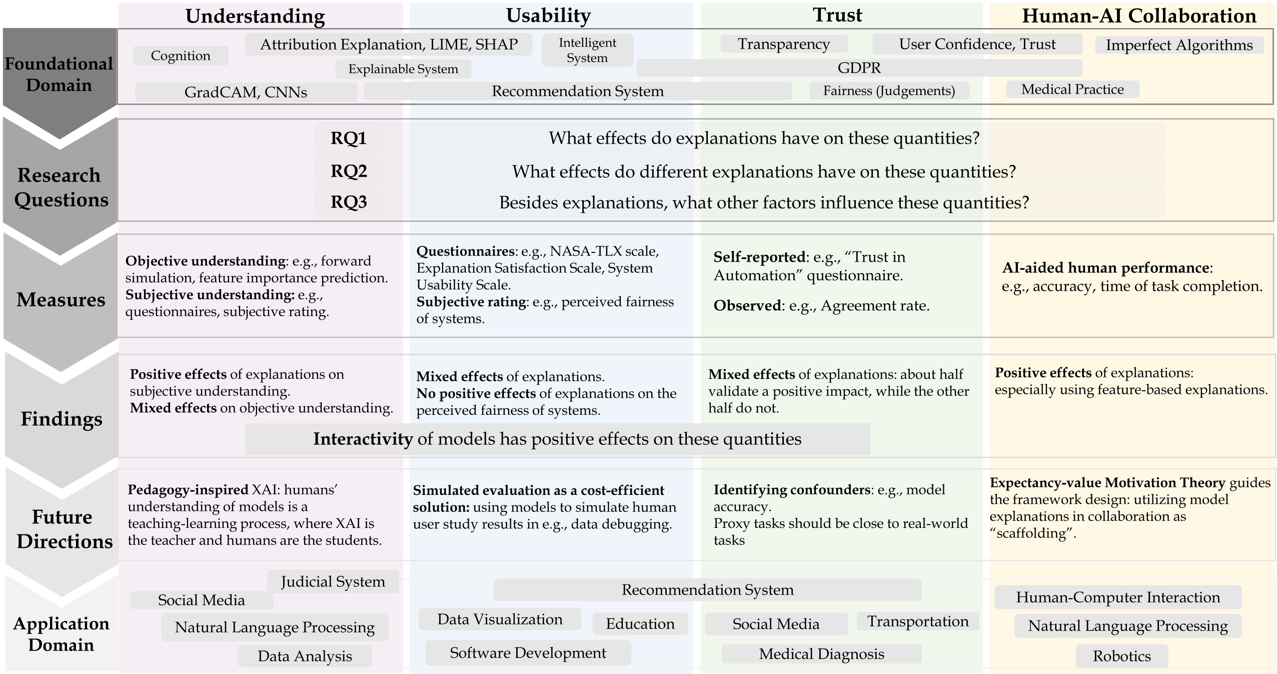
1) To offer an overview of the foundational work of user studies in XAI, we investigated references of all 97 core papers in a data-driven manner. Likewise, we analyzed follow-up works building on these core papers (identified through citations of core papers) to reveal the fields impacted by XAI user evaluations (Section III).
2) We present a summary of the design details in XAI user studies with particular focus on the deployed models and explanation techniques, experimental design patterns, participants as well as concrete measures, providing inspiration of how to collect human assessment (Section IV).
3) We discuss the impact of using explanations on different aspects of user experience (Section V), which can serve as an overview of the effectiveness of the current XAI technology and a summary of the state-of-the-art.
4) Based on the examined user study details and their best-practice findings, we synthesize guidelines for designing an effective user study for XAI (Section VI).
5) Beyond the user study design, we discuss potential paradigms of AI systems understanding humans in the context of e.g., theory of minds, as well as other future research directions (Section VII).

Our study highlights under-investigated areas in the context of current user-centered XAI research such as cognitive or psychological sciences through data-driven bibliometric analysis. Together with our proposed guidelines, we believe that this work will benefit XAI practitioners and researchers from various disciplines and will help to approach the overarching goal of human-centered XAI.

II. Related Work

As a vast amount of explanation methods have been proposed, many researchers seek a systematic overview of the ever-growing field of XAI. In [128], [129], [130], [131], [132], [133], the authors aim to cover many facets of XAI technologies ranging from problem definitions, goals, AI/ML model explanations to evaluation measures, while in [134] the authors emphasize the research trends and challenges in Human-Computer-Interaction (HCI) applications. A large body of XAI surveys focuses mainly on the interpretability of a particular family of models and corresponding explanation techniques. For instance, [135], [136], [137] investigate explanations for Deep Neural Networks (DNNs), where models often take images as input [135], [136]. Joshi et al. [137], however, provide an extensive review for DNNs with multimodal input for instance that of joint vision-language tasks. Causal interpretable models are gaining more attention recently and Moraffah et al. [138] provide a literature review for causal explanations. A systematic literature review on explanations for advice-giving systems is conducted in [139]. Among these surveys focusing on general XAI technologies, evaluation measures are only briefly examined.

One challenge in XAI research is to evaluate and compare different explanation methods, due to the multidisciplinary concepts in interpretability/explainability [119], [120], [140]. Evaluation measures can be divided into two groups: human-grounded measures that rely on human subjects and functionally-grounded metrics that can be computed without human subjects [119], [120]. Many researchers seek solutions to evaluate explanations automatically. A comprehensive literature review with a focus on these functionally-grounded evaluation methods (without human subjects) can be found in [120]. Explainability is an inherently human-centric property, therefore, the research community should and has started to recognize the need for human-centered evaluations when working on XAI [119], [141].



	Understanding	Usability	Trust	Human-AI Collaboration
Foundational Domain	Cognition Attribution Explanation, LIME, SHAP Explainable System GradCAM, CNNs	Intelligent System Recommendation System	Transparency User Confidence, Trust GDPR Fairness (Judgements)	Imperfect Algorithms Medical Practice
Research Questions	RQ1 RQ2 RQ3	What effects do explanations have on these quantities? What effects do different explanations have on these quantities? Besides explanations, what other factors influence these quantities?		
Measures	**Objective understanding**: e.g., forward simulation, feature importance prediction. **Subjective understanding:** e.g., questionnaires, subjective rating.	**Questionnaires**: e.g., NASA-TLX scale, Explanation Satisfaction Scale, System Usability Scale. **Subjective rating**: e.g., perceived fairness of systems.	**Self-reported**: e.g., “Trust in Automation” questionnaire. **Observed**: e.g., Agreement rate.	**AI-aided human performance**: e.g., accuracy, time of task completion.
Findings	**Positive effects** of explanations on subjective understanding. **Mixed effects** on objective understanding.	**Mixed effects** of explanations. **No positive effects** of explanations on the perceived fairness of systems.	**Mixed effects** of explanations: about half validate a positive impact, while the other half do not.	**Positive effects** of explanations: especially using feature-based explanations.
	Interactivity of models has positive effects on these quantities			
Future Directions	**Pedagogy-inspired** XAI: humans' understanding of models is a teaching-learning process, where XAI is the teacher and humans are the students.	**Simulated evaluation as a cost-efficient solution:** using models to simulate human user study results in e.g., data debugging.	**Identifying confounders**: e.g., model accuracy. Proxy tasks should be close to real-world tasks	**Expectancy-value Motivation Theory** guides the framework design: utilizing model explanations in collaboration as “scaffolding”.
Application Domain	Judicial System Social Media Natural Language Processing Data Analysis	Recommendation System Data Visualization Education Software Development	Social Media Transportation Medical Diagnosis	Human-Computer Interaction Natural Language Processing Robotics

Fig. 1. Roadmap of our literature analysis. We find out the foundational works of core papers and their application domains using a data-driven method introduced in Section III. Three main research questions in user studies are distilled from core papers. Methods related to measures of each category are discussed in Section IV, and findings of the research questions are summarized in Section V. Based on the findings, we propose future directions to further promote human-centered XAI in Section VII. We distill important messages in this figure, but refer to the discussion in the corresponding sections for more details.

For instance, Chromik and Schuessler [125] propose a taxonomy on XAI evaluations involving humans. Mohseni et al. [126] summarize four groups of human-related evaluation metrics: mental model (e.g., user's understanding of the model), user trust, human-AI task performance and explanation usefulness and satisfaction (i.e., user experience). Hoffman [124] places more focus on psychometric evaluations by proposing a conceptual model of the XAI process and specifying four key components that should be evaluated: explanation goodness and satisfaction, (user's) mental models, curiosity, trust and performance. Beyond assessing evaluation methods, XAI applications are designed to eventually support decision-making and benefit end users. A recent review by Lai et al. [142] considers studies on collaborative Human-AI decision-making, which may include AI agents providing explanations. Success in human-AI decision-making tasks can be seen as one amongst many other ways to evaluate the effect of explanations. Ferreira and Monteiro [143] present a review of the user experience of XAI applications to answer who uses XAI, why, and in which context (what + when) the explanation is presented.

Closer to our focus on user studies concerning XAI, Liao et al. [141] study user experiences with XAI to reveal pitfalls of existing XAI methods, underscoring the important role of humans in XAI development. As suggested by Doshi-Velez and Kim [119], a human-subject experiment needs to be designed sophisticatedly to reduce confounding factors. In contrast to previous surveys on XAI, we aim to provide XAI researchers and practitioners with a comprehensive overview of the research questions explored in user studies, along with thorough information on experimental design. To this end, we present a practical guideline in user study design, which can be used as a starting point for future exploration of human-centric XAI applications.

III. Methodology

To analyze the collected papers related to user studies on XAI, we first categorize them into four groups based on their objectives. From these studies, we distill three main research questions concerning the effects of model explanations on each objective. We then summarize the methods used in these studies to quantify these objectives. Important findings from the papers are discussed, and we propose future directions based on these findings. Additionally, we examine the foundational works upon which these user studies are based (i.e., their references) and the follow-up papers that cite them, shedding light on the foundational works and emerging trends in human-centered XAI studies. Fig. 1 presents a roadmap of our analysis.

In this section, we first describe the criteria used for their categorization. We then discuss the foundational and application domains of these papers, providing a broader view before diving into their detailed analysis.

A. *Categorization of User-Study Objectives*

Since the core papers cover various factors of model explanations, we decided to categorize the core papers into different clusters to better study their commonalities and differences. In [119], *interpretability* in the context of ML systems is defined as the ability to explain or present model predictions in understandable terms to a human. Beyond fostering comprehension, the authors argue that interpretability can assist in qualitatively ascertaining whether other desiderata, such as *usability* and *trust* are met. During a profound study of the relevant literature that was previously selected, we identified four sensible categories, that are derived from the considered dependent variables in user studies (desiderata of interpretability). These four categories are *trust, understanding, usability*, and *human-AI collaboration performance*. In Table I, the studied papers are categorized according to the measured quantities. As each measure can usually be assigned to only one of these categories, we found this distinction to be intuitive.

These categories reflect different functionalities (goals) of XAI. As interpretability is defined as "*the ability to explain or to present in understandable terms to a human.*", humans' "understanding" is the direct goal of XAI. To be concrete, understanding in the context of interacting with an ML model refers to a user's grasp or "mental model" of how the model operates, and this knowledge grows from using the system and from clear explanations about it [141]. "Usability" is commonly studied in human-computer interaction [144], which is one of the desiderata of XAI [119]. According to [145], usability is the extent to which users can utilize a product to successfully, efficiently, and satisfactorily accomplish their intended objectives. Thus, this category encompasses user studies that employ model explanations to support users in achieving specific tasks. In usability, different aspects are measured, for instance, whether the system is easy to use or how much cognitive load it requires. The aspect "undesired behavior detection" relates to use cases where explanations uncover model discriminatory behaviors, such as the utilization of undesired features. "Trust" in AI is summarized as a combination of the user's confidence in a model's accuracy, a personal comfort level with understanding and using it, and the willingness to let the model make decisions [140]. It encompasses more requirements. Human-AI collaboration performance is related to scenarios where the AI system provides its predictions, but humans retain the final decisions [89]. In this case, model explanations are deployed to reach a performance superior to that of the AI system or the human decision-maker alone. These categories cover different dependent variables of interest in the reviewed user studies, primarily related to how XAI methods function. These functions mainly tie to the models' reasoning and knowledge representation. A wider perspective on XAI, which assesses generalization or robustness, remains an important field for future exploration through user studies.

B. Foundations of User Studies

Based on a data-driven bibliometric analysis of the references in core papers, we highlight significant research topics within the "Foundational Domain" in Fig. 1. It is evident that model explanations and interpretability are pivotal components. This includes papers that introduce explanation methods such as LIME [146], SHAP [147], and other attribution methods. These are a frequent subject of study in works measuring understanding and usability. Additionally, convolutional networks, which are commonly employed in experiments, use tools like GradCAM [148] and various saliency maps to generate model explanations. Notably, many research papers appear within the domain of recommender systems, because many XAI user studies are conducted in the context of recommendation solutions. he EU's General Data Protection Regulation (GDPR) [149] is frequently mentioned in core papers due to the ongoing debate on the right to explanation" [150]. This debate has significantly influenced the shift in modern AI systems towards explainability. While the ultimate consumers of model explanations are humans, well-established research domains that focus on human understanding are underrepresented. For instance, only a few papers related to "Cognition" are cited compared to those on other algorithmic topics. Millecamp et al. [18] suggest enhancing XAI theory with insights from social sciences, including cognitive science and psychology. Given the scant references to psychology, it appears that only a handful of XAI user studies delve into evaluating XAI from a psychological standpoint. We highlight a nascent research domain of XAI frameworks based on human cognition and behavior theories [141]. This theoretical guidance can also offer conceptual tools for better evaluating XAI from user perspectives. More details about common references can be found in Appendix A.1, available online.

C. Impact of User Studies

Fig. 1 presents applications that make use (and thus are the consumers) of the findings from core papers. We noticed that studies on user understanding and trust span a wide range of applications. For example, trust is frequently addressed in the contexts of medical diagnosis and transportation, indicating its significance in high-risk scenarios. Recommendation systems emerge as a primary focus in follow-up works. Papers on usability have a significant impact on fields like data visualization, software development, and education. In these areas, models frequently serve as tools to ease the burden on end users. Human-AI collaboration measures particularly promote the further development of robotics and or natural language processing. The prominence of recommendation systems in both foundational works and their impact implies that XAI is an integral component of contemporary recommendation systems. A comprehensive overview of the fundamental works and application domains can be found in Appendix A.1, available online.

IV. Comprehensive User Study Analysis

In this section, we present details of the covered XAI user studies. We first introduce some commonly used AI models and explanation techniques (Section IV-A), followed by a discussion of application domains and measures with respect to the four measured quantities. The experimental designs, as well as analysis tools are presented in Section IV-C.

TABLE II
KEYWORDS FOR OUR PAPER SEARCH QUERY

	Explainable AI	User Study
Keywords	XAI, explainable AI, explanation, explainable, explanatory, interpretable, intelligible, black-box, machine learning, explainability, interpretability, intelligibility, explain attribution, feature	user study, participant , human subject, empirical study, lab study, user evaluation, human evaluation

Two groups of keywords were used.

TABLE III
MODELS AND EXPLANATIONS IN CORE PAPERS

		White-box	Black-box	Other
Feature-based	**local**	[21, 48, 153] [12, 22, 39] [6, 50]	[21, 45, 49, 55] [29, 34, 72, 92] [35, 39, 40] [42, 47, 65] [54, 57, 58] [50, 56, 71] [40, 41, 89] [25, 59, 90] [43, 60, 95]	
	global	[12, 53, 74] [21, 50]	[50]	
Example-based		[12, 21, 43] [6, 74, 96]	[17, 52, 57] [13, 25, 40]	[32] (generative models)
Counterfactual		[12, 37] [21, 82]	[27, 100] [57, 65]	
Concept-based			[61, 62, 71] [63, 64, 67] [57, 99]	
Other		[11, 88] [7, 10]	[1, 9, 15, 154] [3, 13, 51] [3, 56, 58] [36, 49, 55] [16, 28, 85] [33, 38, 68]* [8, 23, 76] * [69, 70]*	[2] [18, 19, 20] † [20, 26, 84] † [66, 81, 83] † [14, 30, 85] † [5, 91] †

Papers are categorized according to types of explanations (column) and types of models (row). ∗ denotes papers using recommendation systems as models; † denotes papers proposing novel interpretable interfaces as studied models.

A. *Models and Explanations*

As our selected core papers comprise a large spectrum of AI models, data modalities, and explanation approaches, we initially list the models and explanation techniques deployed along with the corresponding core paper references in Table III. It presents the utilization of explanation types in columns and model types in rows. The explanation methods used is organized according the the taxonomy by Molnar [151]. First, there are intrinsically interpretable models, also known as *white-box models*. For instance, white-box models include decision trees and linear models. Second, there are *black-box models* that provide no parameter access or are too complex to be explained in a human-understandable way [152]. These include ensembling techniques such as Random Forests or neural models.

As for explanation techniques, we identified five key types in the scope of the surveyed papers (rows of Table III). Most frequently used are feature-based (attribution) explanations, for instance, SHAP (Shapley additive explanations [147]) and LIME (Local Interpretable Model-Agnostic Explanations [146]). There is a clear differentiation between local, instance-wise, explanations and global explanations that apply to the model in its entirety. For instance, the weights of a linear model have a global scope. This differentiation is common among these feature-based explanations, where most of the papers using local explanations. Other popular explanation types are example-based explanations, counterfactual explanations, which aim at providing actionable suggestions for attaining a user-preferred prediction by changing certain input features, and concept-based explanations, which use meaningful high-level concepts such as objects or shapes to explain a prediction.

Besides these four main types of explanations, there are other explanations such as rules [11], [88] or game strategies [7], [10] when AI plays games. More details about concrete models and explanations can be found in Appendix B, available online.

B. *Measurements*

The effectiveness of explanations can be characterized from several angles. We specifically identified the categories of trust, understanding, usability, and human-AI collaboration performance. In this section, we give an overview of the contexts in which each of these variables is studied and the measures used to quantify them.

1) Trust: User trust is studied in decision-making applications such as image classification [13], [17], (review) deception detection [25] or loan approval [27]. Besides decision making, [5], [8], [16], [18], [19], [23] study user trust in the domain of recommendation systems. Whether explainable ML models can increase user trust in the medical domain is studied in [1], [6], [9]. Moreover, Colley et al. [3] measure user trust in an autonomous driving application with and without explanations.

Trust measures used in much of the existing research can be divided into two groups: *self-reported* and *observed* trust [155]. Self-reported trust is commonly measured by asking users to fill out questionnaires whereas observed trust is quantified by humans' agreement with the model's decisions. In Table III in Appendix, available online, trust measures in these two groups are listed. The agreement rate of users with the model decisions is commonly used [9], [11], [12], [25] as a measure of observed trust. Parallel to observed trust measurement, van der Waa et al. [156] ascribe the user's alignment behaviors to the *persuasive power* of model explanations, i.e., the capacity to convince users to follow model decisions despite the correctness. As an extension, trust calibration is defined based on this measure. For example, a high agreement rate to wrongly made decisions represents *overtrust*, while a low agreement rate to correct decisions means *undertrust* [12]. In self-reported measurements, researchers either utilize well-developed questionnaires or self-designed ones, with the exception of [4] which conducts a semi-structured interview to explore user opinions. Several works [6], [11], [13], [16], [17], [18], [19], [24], [27] propose their own questionnaires. Among these, a subgroup [13], [16], [18], [19], [24] simply asks users to rate a single statement such as "I trust the system's recommendation/decision", which is named as one-dimensional trust by [8]. When deploying previously proposed questionnaires [2], [3], [5], [7], [8], [10], [21], [22], [23], [157], Trust in Automation [158] is the most commonly used one, in which the underlying constructs of trust between human and computerized systems are explored.

2) Understanding: An important goal of explanation techniques is to foster users' understanding of complex ML systems. An important separation has to be made between users' perceived understanding and their actual comprehension of the underlying model, as the two often do not agree [35], [40]. Cheng et al. [22] explicitly differentiate between *objective* understanding and self-reported understanding, which we term *subjective* understanding in this work. While subjective understanding is usually measured through questionnaires, measuring objective understanding requires a proxy task where the users' understanding is put to a test. Additionally, user studies can be run to assess how well users can understand the explanation itself (and not the underlying model). This can be an important sanity check and is particularly used in the domain of conceptual explanations [62], [159], where the intelligibility of concepts needs to be verified. We refer to the third category as *understanding of explanations* but defer its detailed findings to Appendix C.3, available online.

Objective Understanding: Works in the subdomain of objective understanding deploy proxy tasks to verify users' understanding of a model's inner workings. The most commonly considered domain in works on understanding is finance [35], [39], [40], [47], [48], [49], [53] followed by image classification [13], [21], [52]. One of the most critical design choices when assessing objective understanding is the selection of a suitable proxy task. Doshi-Velez and Kim [119] argue that the task should *"maintain the essence of the target application"* that is anticipated. One of the most prominent tasks is forward simulation [119], [140]. This task demands subjects that are given an input to simulate, i.e., predict, the model's output. The extent to which participants can successfully provide the model's output is also referred to as *simulatability* [140]. However, scholars have designed many more tasks to quantify understanding and applied them across a variety of data modalities (cf. Table 2 in Appendix, available online for an exhaustive listing).

We briefly describe other common tasks below. A special variant of forward simulation is called *relative simulation*. In this task, users predict which example out of a predefined choice will have the highest prediction score (or class probability). *A manipulation or counterfactual simulation task* [119] asks users to manipulate the input features in such a way that a certain model outcome (counterfactual) is reached. Users' performance on this task can be used as a proxy for their understanding. Lipton [140] pointed out that simulatability can only be a reasonable measure, if the model is simple enough to be captured by humans and that simpler tasks are required otherwise. An example could be a *feature importance* query, where users have to tell which features are actually used by the model. A directed and more local version of this task is *marginal effects queries*, where the subjects predict how changes in a given input feature will affect the prediction (e.g., *"Does increasing feature X lead to a higher prediction of Y being class 1?"*). Because explanations should allow the identification of weaknesses in models, the task of *failure prediction* measures the accuracy of users' prediction when the model prediction is wrong.

Subjective Understanding: Besides the objective understanding which is supported by performance indicators, understanding of a model may be subjective, i.e., it may depend on a user's own perception. The most commonly used applications that measure subjective understanding are various recommendation system setups [16], [33], [34], [38].

Most of the works assess the subjective understanding of a user with a post-task questionnaire. Guo et al. [7] adapted a popular questionnaire designed for recommendation systems by Knijnenburg et al. [160], while Bell et al. [39] accommodated the questionnaire which originally intended to measure the intelligibility of differenet explanations by Lim and Dey [161]. On the other hand, agreement to simple subjective statements such as *"I understand this decision algorithm"* [22], *"I understand how the AI..."* [13], [17] or *"The explanation(s) help me to understand..."* [33] can be collected to assess subjective understanding.

3) Usability: Usability is a key concern of every HCI system and thus applies to almost all domains. This is reflected in the surveyed papers, where usability is studied in a wide range of setups and contexts. We also include application-specific performance measures in this category.

Based on the measurements in the user studies, we refined usability into measures of helpfulness, workload (cognitive load), satisfaction, ease of use and detecting undesired behaviors of the system, as shown in Table I. To assess workload (cognitive load), NASA-TLX scale [162] is used in [3], [6], [16], [21], [66], while Abdul et al. [48] measure cognitive load by capturing the log-reading time of memorizing the explanation. Most of the works use self-designed questionnaires or statements to measure satisfaction [6], [16], [18], [19], [29], [30], [69], [70], however, the Explanation Satisfaction Scale [163] can be deployed as an established alternative [1], [47]. Helpfulness can be assessed by simply asking for subjective ratings of the explanations for accomplishing a specific task [13], [46], [56], [65], [67], [68]. Colley et al. [3] use an adapted version of the System Usability Scale proposed in [164].

Using model explanations to audit models is one purpose of explainability [129]. Some of the surveyed works study how model explanations can assist users in detecting undesired behaviors of models. These issues mainly include (perceived) unfairness in the model decision-making [38], [74], [78], [79], biases in models [72] or features [57], and wrong decisions (failures) [24] in the studied papers. A detailed summary of types of undesired behaviors is listed in Table VI. In the undesired behavior detection, the effectiveness of explanations is evaluated by objective performance measures, such as the number of bugs identified [71], the share of participants that identify a certain bias [57, First Experiment] or by the deviations between model predictions and human predictions for unusual samples [53]. The perception of users regarding fair treatment by a system has primarily been researched in high-stakes applications such as granting loans [27] or granting bail for criminal offenders [73], [74], [75]. For example, [73], [74], [75] investigate the fairness of COMPAS, a commercial criminal risk estimation tool that was

TABLE IV
EXPERIMENTAL DESIGNS IN CORE PAPERS

	Experimental Design		
	Between-Subjects	Within-Subjects	Mixed
	[5, 7, 8, 12, 15, 27, 59]		
	[17, 21, 22, 23, 25, 72]	[1, 3, 4, 9, 19]	
	[11, 28, 32, 40, 46, 95]	[10, 18, 21, 24, 70]	[2, 13, 16, 52, 66]
Papers	[47, 49, 50, 51, 53, 84]	[13, 26, 35, 52, 91]	[10, 28, 34, 64, 74]
	[36, 37, 38, 39, 43, 56]	[57, 71, 78, 81, 93]	[12, 33, 65, 83, 146]
	[29, 30, 54, 90, 92, 96]	[6, 62, 63, 67, 69]	[61]
	[12, 38, 57, 58, 73]	[14, 41, 45, 60, 68]	
	[75, 76, 77, 78, 82]		

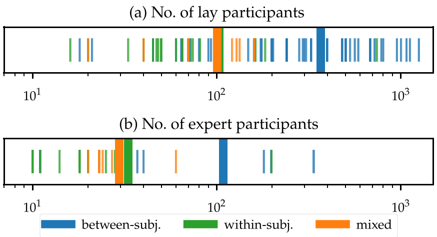


Fig. 2. Distribution of participant numbers in the surveyed user studies by design and participant type (each bar represents one study). Per-design means are indicated in bold.

used in the US to help make judicial bail decisions. It is also considered in everyday use-cases such as news [38] and music [77] recommendations, or possible career suggestions [76], where a bias in the underlying system can be to the detriment of the user. As the assessment of fairness is a very subjective matter, questions regarding perceived fairness are prevalent, e.g., “how the software made the prediction was fair” [74], which can be answered on 5- or 7-point Likert scales [2], [27], [38], [73], [74], [75]. Among these works, an effective explanation is the one that can either increase or decrease the fairness perceptions, since the aim of explanations is to show fairness or unfairness. An exhaustive overview of measures for usability is given in Table IV of the Appendix, available online.

4) Human-AI Collaboration Performance: The goal of human-AI teaming is to improve the performance in AI-supported decision-making above the bar set by humans or an AI alone [89]. Improving human performance with the help of AI has been considered in games [10], [88], question answering tasks [89], [91], deception detection [25], [90] and topic modeling [29], [30].

The most common assessment is to rate AI-aided human performance by the percentage of correctly predicted instances in the decision-making process [25], [89], [90]. Paleja et al. [10], however, define the performance as the time to complete the task. In [88], performance is measured in a game-based application, chess, using a winning percentage (which is commonly used in sports) as well as a percentile rank of player moves.

C. Experimental Design and Analysis

There are three common experimental settings when conducting user evaluation: between-subjects (or between-groups) designs, within-subjects designs, and mixed designs that combine elements of both. An overview of the designs found in the core papers and their participant numbers is presented in Table IV and Fig. 2, respectively.

1) Between-Subjects: With slightly above 55 % of the user studies conducted in a between-subjects manner, i.e., one subject is only exposed to one condition, this design choice is most common in the XAI literature. The number of participants in the between-subjects manner usually starts at around 30 participants, while it may go up to 1070 in total for 3 conditions as in [17] and to 1250 for 5 conditions in [53]. However, the number of participants can be limited when the studied application is designed for specific groups of lay persons, which cannot be easily recruited from the Internet platforms such as Amazon Mechanical Turk. For instance, Ooge et al. [8] use 12 school students per condition. Some authors place particular emphasis on participants being similar to the average demographic [73], [75].

The conditions usually include the different explanation techniques in combination with other parameters such as the model, data set, data modality, or a number of features used as independent variables. Note that a full grid design with many independent variables may quickly result in a very high number of conditions, which in turn requires many participants. The outcome variable of interest is commonly measured on a numerical or ordinal scale right away, however, in the fairness domain, qualitative analyses are sometimes obtained through conducted interviews or written responses [2], [27], [73].

The statistical analysis directly follows from this design. If one is interested in identifying significant differences between the groups, common statistical hypotheses tests are used. For overall comparison, one or two-way ANOVA tests are the most commonly used statistical tool. Interesting post-hoc comparisons between two groups can be made with a standard T-test, if the data is normally distributed with equal variance, or by using non-parametric tests such as the Wilcoxon rank-sum test (also known as Mann-Whitney U-test) for comparison of two populations (e.g, [57]) or the Tukey HSD test (e.g., [49]) for multiple populations. When running multiple post-hoc tests, some works make use of the Bonferroni correction (e.g, [57]).

2) Within-Subjects: Around 30 % of the papers use the within-subjects design, where each participant sequentially passes through all conditions and provides feedback. Fewer participants are recruited in within-subjects experiments compared to the between-subjects ones. Hence, they are particularly popular when participants with restrictive characteristics, such as domain-specific professional expertise, are required. For example, Suresh et al. [9] and Rong et al. [26] recruit fourteen medical professionals and five radiologists in their user studies, respectively. The small number of medical experts contributing to the user study is a limitation [26], however, it is often the case in expert user research. Gegenfurtner et al. [165] evaluate 73 sources and point out that the majority of these studies include only five, maybe ten experts. Besides the medical domain, other works [3], [4], [19], [21] also invite subjects with particular professions such as engineers in a technology company. When no specific knowledge is required, however, participant

TABLE V
USER STUDY FINDINGS WHEN USING MODEL **EXPLANATIONS** AS EVALUATION DIMENSIONS

		Evaluation Dimension: Explanations Effect of explanations compared to **no explanations**	
		Positive	Non-positive / Mixed
Trust		[13]: example-based, rule-based explanations [16]: example-based explanations for recommendations [27]: feature importance [10]: decision-tree explanation for policy [28]: explanation corpus given by researchers [25]: feature-based (saliency map), example-based explanations [8]: explanations for recommendations [6]: rationale-based, example-based and feature-based (best) explanations for online symptom checkers [15]: confidence scores	[3]: positive in simulation but no improvement in real-word [1]: explanations for medical suggestions (Doctor XAI [155]) pos.for observed trust but insignificant for reported trust [12]: feature-based explanations increase appropriate trust slightly but counterfactual explanations inconclusively [21, 22, 24]: feature-based explanation, insignificant [11]: rule-based explanation, insignificant [15]: Shapley values, insignificant [29]: feature-based explanation, negative
Understanding	Obj.	[22, 53] white-box model [40] feature importance, LIME (tabular) [46] counterfactuals+cues (audio) [50] manipulatability improved by white-box log. reg. [54] saliency maps (image) [59] saliency maps for bias detection and strategy identification [12] counterfactuals+feature importance	[39]: SHAP, negative for black-box model (education domain) [39]: Insignificant difference btw. black-box and white-box models [40]: Prototypes, Anchors, LIME on textual data insignificant [46]: Counterfactuals and Concepts insignificant (audio data) [50]: Simulatability results insignificant for LIME, IG, surrogate model on BERT and Logistic Regression Model, Manipulatability insignificant for BERT [58]: Insignificant results for GRAD-CAM, Saliency Map, uncertainty scores in VQA [59] saliency maps for failure prediction (image) [60]: saliency maps, negative for a mix of three interpretation techniques in simulation task
	Sub.	[13]: example-based, rule-based explanations [28]: explanation corpus given by researchers [12]: feature-, example- and counterfactual-based [38]: explanations provided by [167] for Facebook News Feed [16]: example- and feature-based explanations [17]: example-based explanations [34]: feature importance, SHAP and LIME [35]: feature importance, SHAP	[22]: white-box model, insignificant [39]: white-box < black-box, both insignificant [31]: feature importance explanation (transparent system) can be distracting
Usability		[81]: counterfactuals, pos. for usability [16, 47]: example-based explanations, pos. for satisfaction [67]: CAM-related explanations, pos. for helpfulness [6]: rational-, feature-, example-based explanations, pos. for satisfaction [70]: content-based explanations, pos. for satisfaction [83]: explanations regarding driving information, pos. for ease of use [13]: example-based and rule-based explanations, pos. for helpfulness [71]: local, global, visual (saliency map) explanations, pos. for bug identification [65]: attribution methods and conceptual explanations, pos. for usefulness [84]: feature-based, pos. for reliability [24]: (proposed) template-based expl. pos. for debugging and usefulness [27]: feature importance, counterfactual explanations pos. for perceived fairness	[82]: counterfactuals, significant for helpfulness/usability but insignificant for usefulness [1]: ontology-based explanation, insignificant for satisfaction [65]: attribution methods and conceptual explanations, insignificant for ease of use [24]: visual explanations increases usefulness, but improvement is insignificant [3]: pos. for cognitive load/usability (simulation), but insignificant in real-world [29]: feature-based explanations, negative for satisfaction [38]: informing users about the algorithmic decisions, negative ranking scores of recommendations, insignificant for perceived fairness [27]: highlight features only, insignificant for perceived fairness [78]: insignificant in between-subjects but significant in within-subjects for perceived fairness
Human-AI Collaboration Performance		[88]: textual explanations with domain knowledge (in chess) [25, 90]: feature-based explanations [91]: exampled-based for experts, feature-based for novices [93]: contrastive explanations [13]: example-based and rule-based explanations [95, 96]: example-based explanations, attributions (AI correctness prediction) [96]: important parts in images as explanations	[25]: exampled-based, insignificant [15, 89]: feature-based explanations, insignificant

Effects of explanations compared to the baseline (control group) of “no explanations” on measured quantities. Effects are divided into “positive” where explanation information is given, and “non-positive / mixed” where negative impact is marked with underlines.

numbers reach up to 740 also for within-subjects designs [93]. For within-groups designs, the Wilcoxon signed-rank test (e.g., used by [35], [52]) is the most common method to compare paired samples for significant differences. Repeated-measures ANOVA is a common analysis tool, when multiple comparisons are required (see, e.g., [35]).

3) Mixed: The smallest group of studies, about 15%, use a mixture of between- and within-subjects settings. In these works, subjects are first assigned randomly to one group, where they are exposed to multiple conditions. Anik and Bunt [2] use knowledge background in machine learning as a between-subjects factor to divide the participants into three groups (expert, intermediate and beginner), while inside each group participants interact with explanations in the context of four different scenarios (e.g., facial expression recognition or automated speech recognition). Dominguez et al. [16] make the presence of explanations a between-subjects condition and different types of explanations a within-subjects factor in the group with model explanations. A particular challenge for such a study design is that statistical tools from both the independent-samples and dependent-samples categories need to be combined.

V. FINDINGS OF USER STUDIES

In this section, we summarize the primary findings from the core papers. Table V lists findings with respect to four measured quantities. To build an overview of the findings, we divide papers according to their evaluation dimensions, i.e., the independent variables in the user studies. When using the presence of explanations as the evaluation aspect, the findings are summarized in Table V. The listed impacts using explanations are to be seen in comparison with a control group without explanations. Effects are divided into two groups: (1) Positive effects, for example, increasing user trust or understanding; (2) Non-positive effects: the effect can be negative, or not significantly positive (neural), or a mixture of different effects (e.g., feature-based explanations have positive effects but counterfactual explanations do not). Beyond the explanations themselves, other possible evaluation dimensions such as that might have an impact on the perception of XAI, for instance, AI technology literacy, model performance, or the dimensionality of the data. Instead of using the mere presence of explanations, many works compare different explanation techniques with each other (see Appendix D, available online for more details).

TABLE VI
OVERVIEW OF RESULTS FOR UNDESIRED BEHAVIOR DETECTION USING MODEL EXPLANATIONS

Type	Paper	Detection Result
Wrong decisions (failures)	[24, 53, 71]	High detection rate in [24]; Moderate detection rate (50%) in [71]; Lower detection rate in [53]
Biases in features used by models	[57, 71]	Moderate detection rate (50%) in [71]; Moderately high detection rate (>50%)
Discrimination/Biases in decisions	[72]	Humans perform well in bias detection (accuracy=88.9%) and bias description (66.7%)
Unfairness (perceived) in models	[2, 27, 38, 74] [73, 78, 79]	Succeed to judge [27, 74, 78]; Not succeed to judge [2]; Not always (no consensus) [38, 73]

As various research questions and findings are addressed in 97 core papers, many papers compare explanation types in order to choose a preferable one, it is not possible to cover all results in one table. Based on them, we outline some interesting trends in the effectiveness of explanations on user experience: (1) Explanations are effective in improving users' subjective understanding; (2) The effectiveness of explanations in increasing user trust and usability of models is not clear; (3) Explanations are not good at convincing users that models are fair; (4) Interactivity of the model has positive impact on user trust, understanding and model usability. The first three statements can validated through the number of papers obtaining positive or non-positive effects in each category, while the last finding is extracted from Table V in the Appendix, available online, which details findings with on other independent variables. We encourage the reader to consider the short summary of *primary* findings in the tables and check for further details according to their specific interests. In the following section, we highlight some findings for each category of measurement.

Trust: Among the papers comparing the effect of using explanations to using no explanations, or placebo (randomly generated) explanations [8], [25], about half of the papers validate that explanations have a positive impact on user trust [1], [8], [10], [13], [16], [25], [27], [28], while the other half cannot verify this hypothesis [3], [11], [12], [21], [22], [24]. For instance, Colley et al. [3] investigated the explanations in an autonomous driving task and discover that the trust is improved in simulation but not with the real-world footage. Another example of the mixed effect of using explanations is found in [12], where (minimal) evidence is found that feature-based explanations help increase appropriate trust, but counterfactual explanations do not.

Apart from using explanations as independent variables, the user personalities or expertise may also affect their perceptions [2], [17], [18], [22], [23], [30]. Millecamp et al. [18] captured personal characteristics in the aspects such as the Locus of Control defined by Fourier ("the extent to which people believe they have power over events in their lives"), Need for Cognition ("a measure of the tendency for an individual to engage in effortful cognitive activities") or Tech-Savviness ("the confidence in trying out new technology"). However, no significant interaction effect could be found between the personal characteristics and the trust. Liao and Sundar [5] studied a recommendation system asking users' personal data with different explanations. They hypothesized that explanations in a "help-seeker" style and using the pronoun "I" would gain more trust of users than the explanations formalized in a "help-provider" style. Nevertheless, However, the opposite result is found and using self-referential expression resulted in lower affective trust. Model performance together with model explanation was studied in [17] for an image recognition task. The authors found out when images were recognized (high model performance), users feel the system more capable ("capability" is defined as a belief of trust).

Understanding: The fundamental question in this subdomain is to find out which explanation technique is most beneficial for increasing the user's understanding of a machine learning model. As pointed out earlier, understanding can be measured both in a subjective and objective manner.

We first discuss results on objective understanding. The goal of increasing objective understanding was explicitly posed by Alqaraawi et al. [54] who reported that saliency maps have a positive effect on understanding. Wang and Yin [12] show that counterfactual explanations and feature importance increase users objective understanding. On the contrary, Sixt et al. [57] find none of their examined explanation techniques (counterfactuals, conceptual explanations) superior to a baseline technique consisting of example images for each class and the work by Hase and Bansal [40] reveals that many explanations (including anchors, prototypes) have no effect in increasing objective understanding, which LIME on tabular data being the only exception. Apart from the explanation, several other factors have been identified to have an effect on objective understanding. Hase and Bansal [40] suggest that the *data modality* may have a non-negligible impact on how different explanation techniques increase understanding. Some results highlight that the *choice of proxy task* is influential. Arora et al. [50] show that their manipulatablity task revealed differences remained hidden when forward simulation is used. In spite of these findings, Buçinca et al. [13] underline that preferred explanations may be different in a real-world application from a simulated one. Regarding the *type of model*, there is disagreement on whether white or black-box models can lead to increased objective understanding. While black-box models without explanations resulted in higher simulation performance than white-box models with SHAP values in [39], Cheng et al. [22] observe that white-box models increase simulatability and also conclude that *interactivity* is an important factor when it comes to objective understanding.

In comparison with the objective understanding, the research question in the subdomain subjective understanding is to find out how explanations impact user's *perceived* understanding [7], [12], [17], [22], [32], [33], [34], [37], [56]. There exist a trend of using model explanations to improve subjective understanding [13], [16], [17], [28], [34], [38], [167]. However, Chromik et al. [35] challenge the improvement in perceived understanding with the cognitive bias named *illusion of explanatory depth* (IOED) [168], which means that laypeople often have overconfidence bias in their understanding of complex systems. Their results confirm the IOED issue in XAI, i.e., questioning users' understanding by asking them to apply their understanding in practice consistently reduces their subjective understanding. Explanations can have different impacts on subjective and objective understandings [22], where white-box explanations increase objective understanding but do not have significant impact on subjective understanding. Similar disagreements have been observed in multiple other works [40], [167]. Radensky et al. [33] examine the joint effects of local and global explanations in a recommendation system and their results provide evidence that both are better than either alone.

Usability: Similar to trust, it is not clear whether explanations are effective in improving users' perceptions of helpfulness, satisfaction or other dimensions of usability. For instance, in [16], [30], [47], the explanations have a positive effect on satisfaction, while no significant effects on satisfaction are observed in [18], [19], [29], [69]. Parallel to trust, Smith-Renner et al. [29] provide evidence for the hypothesis that it is harmful to user trust and satisfaction to show explanations by highlighting the important words in a text classification task. A strong correlation between self-reported trust and satisfaction can also be observed in [3], where explanations have a positive impact in a simulated driving environment, but no significant effects when using real-world data. Beyond explanations, Nourani et al. [56] study the order of observing system weakness and strengths, which reveals that encountering weakness first results in a lower rate of usage of system explanations than encountering strength first. Schoeffer et al. [27] find out that showing feature importance scores or counterfactual explanations (or a combination of both) for explaining decisions helps increase the perceived fairness, whereas highlighting important features without scores does not. However, several studies don't show a significant difference between scenarios with and without explanations [27], [38], [78]. Effects of explanations may be dependent on input samples, as shown in [67]. The authors show that both Debiased-CAM and Biased-CAM improve the helpfulness for a weakly blurred image, however, there is no significant improvement for unblurred or strongly blurred images. When used to assist users in detecting undesired behaviors, model explanations are likely to identify various types of problems that exist within models or data, as demonstrated by [57], [71], [72]. However, successful detection is not guaranteed. For example, Poursabzi-Sangdeh et al. [53] show that users with model explanations are less able to identify incorrect predictions. A limitation of current detection methods is that users may have varying assessments, such as perceived unfairness and irrelevance [53], [71], [73], regarding the features used in models for decision-making. Due to this limitation, the effectiveness of methods assessed through self-reported data may face challenges in generalizability as discussed in [73]. Yet, these methods generally offer a *one-size-fits-all* solution, failing to account for variations in individual assessments.

Human-AI Collaboration Performance: A strain of works [25], [88], [90], [91], [95], [96], [96] show that viewing explanations can improve human accuracy in making decisions, especially with feature-based explanations taking text data as input [25], [90], [91]. When using example-based explanations in text classification, there is no improvement in human performance [25]. Likewise, utilizing explanations has no significant impact on human performance in [89], [92], but simply showing model predictions has a positive effect in [92]. Experts and novices perceive explanations differently, for example, Feng and Boyd-Graber [91] conclude that the performance gain of novices and experts comes from different explanation sources. Paleja et al. [10] reveal that explanations can improve novices' performance but decrease experts' performance. Additionally, less complex models with explanations can better convince humans in correct decisions [90].

VI. A Guideline for XAI User Study Design

Learning from the best practices of the previous works, we summarize a handy guideline for XAI user study, which serves as a checklist for XAI practitioners. This guideline contains suggestions to avoid pitfalls that researchers could easily overlook. We introduce our guidelines in the order of before, during and after user studies, which reflects user study design, execution and data analysis, respectively.

Before the User Study: When designing a user study, the first step is to decide what to measure. To define the measured quantities, one can consider two alternatives: using a general definition or an application-based quantity that is specific to the application at hand. The former one refers to a quantity that is borrowed from previous well-established research, such as using "trust in automation" [2], [3], [21] or "general trust in technology" [7], [23]. To further construct "trust" as a quantitative measurement, one needs to examine how existing work has conceptualized "trust" in both social sciences context as well as XAI and technical context [169]. The application-based quantity depends on the application goal, for instance in a chess game [88], the measurement is the human winning percentage with the help of model explanations (Human-AI collaboration).

From Table V, we can see that previous works have frequently struggled to prove the effectiveness of XAI even with respect to a control group that is without explanation. When only different explanation techniques are considered, there will always be one winner explanation, but the overall benefit will remain undisclosed (see examples in Appendix D, available online). Therefore, it is important to compare with a baseline without explanations to rigorously show the strength of XAI. When a comparative design is explicitly desired, baselines such as random explanations [28], [41], [62]).

When deploying a proxy task, its difficulty should be gauged and monitored carefully. In the past, the forward simulation task has been criticized as being unrealistically complex for domains such as computer vision [54]. Thus, other proxy tasks such as feature importance queries [57] or manipulatability checks [32], [50] were proposed. Another important point is to choose a proxy task that is simplified, but features many characteristics of the application in mind [119]. Notably, the proxy task should be designed close to the final anticipated application, as even slight differences in the tasks may void the validity of the findings on the proxy tasks in the real world [13].

The measurement is often dependent on the definition of the measured quantity. For instance, in [58], the objective understanding is measured as failure prediction (the accuracy of user prediction when the model prediction is wrong). For subjective measurements such as subjective understanding or trust, one-dimensional measures (i.e., simply rating one question such as *"Do you trust the model explanation?"*) have the drawback that they cannot completely reflect different constructs of measured quantities [8]. Moreover, subjective questions and behavioral measurements often appear to be weakly correlated. For example, the users state that they trust model but they do not really follow the model suggestions [11]. Similar findings have been made with respect to objective and subjective understanding [12], [35], [40]. To overcome this limitation, both self-reported and observed measures shall be used in parallel.

Besides the measures introduced in Section IV-B, there are several psychological constructs that can be deployed to evaluate multiple facets of the interaction between humans and XAI. For instance, the *subjective task value* in the expectancy-value framework is often used to analyze subjective motivation to take any actions [170], which is not thoroughly studied in the XAI experience yet. The subjective task value consists of intrinsic value (enjoyment), attainment value (importance for one's self), utility value (usefulness), and cost (the amount of effort or time needed) [170], [171]. A good explanation interface should be positively correlated with the subjective task value, consequently boosting one's interest and motivation to use the model explanation. With regard to the cost of using model explanations, cognitive load is popularly measured in the current literature with conventional Likert scales [162], [172]. Cognitive load researchers study the validity of different visual appearances in rating scales beyond numerical Likert scales, i.e., pictorial scales such as emoticons (faces with different emotions), or embodied pictures of different weights [173]. Their results demonstrate that numerical scales are more proper in complex tasks while pictorial scales are for simple ones.

Pre-registration using online platforms such as AsPredicted[1] has become a common practice in recent years [174]. In this process, researchers submit a document detailing their planned study online before initiating the data collection. Among other details, the pre-registration includes the measured variables and hypotheses, data exclusion criteria, and the number of samples that will be collected. An exhaustive pre-registration can provide evidence against the findings being a result of selective reporting or p-hacking [175] and thus strengthen the credibility of a study. Expert interviews and pre-studies following a think-aloud protocol [176], e.g., in the references [32], [46], are often mentioned as helpful tools to develop the explanation system and the study design and gain first qualitative insights or complement the qualitative analysis [13], [65].

When preparing for a user study, it is important to plan for explicit steps and to have a backup plan for different situations. Before participants arrive, it is helpful to provide them with information such as where the researchers will meet with them, what they need to bring, and how they can prepare for the study. If conducting the experiment in person, send participants a reminder the day before and provide them with your contact in case they cannot find the experiment site or they need to cancel the experiment session. Once participants arrive, make sure the researchers have a plan that covers all stages of the experiment. The protocol should cover small details (e.g., where participants should leave their backpacks, water bottles, and lunch boxes) and plans for unexpected situations (e.g., uncooperative participants and multifunctional systems). How to obtain participants' consent should be an important part of the procedure. Additional procedure is required for obtaining consent when working with vulnerable populations (e.g., children and pregnant women), in which case alternative consent procedures might take place. Another benefit of pre-designing the experiment script is to fine-tune the language to avoid inadvertent cues. Researchers can unintentionally pass on their expectations to participants through verbal and nonverbal behavior, which might result in participants' skewed performance towards the researchers' desire [169]. To ensure a sound experiment procedure and to protect the integrity of the data, it is worthwhile to put in much effort to design a detailed experiment script.

During the User Study: A sufficient number of participants is the prerequisite of a solid user study analysis. To get a rough estimate of common sample sizes, we refer the reader to the participant statistics in Fig. 2 where we analyze the subject numbers in different experimental designs. For instance, around 350 users without any specific expertise are averagely recruited in between-subject experiments. However, we would like to underline that the required number of participants is highly specific to the study design and should be determined individually, for instance by conducting a statistical power analysis [177]. Additionally, recruited participants should have the same knowledge background as the end users that applications are designed for. For instance, when evaluating an interface explaining loan approval decisions to bank customers, it is not proper to include only students whose major is computer science, since they may have prior knowledge of how model explanations work. Note that the design of an AI application requires different audiences across the project cycle, thus model explanations need to evolve as well [178].

To uphold high-quality standards of the collected data, attention or manipulation checks are essential to filter out careless feedback. This particularly applies to long surveys or online surveys with lay users. Kung et al. [179] justify the use of these checks without compromising scale validity. In within-subject experiments, a random order of conditions is necessary to avoid order effect [1]. Participants can learn knowledge of data or examples shown in the previous conditions, and Tsai et al. [6] choose to use a Latin square design to avoid the learning effect.

After the User Study: After the data collection, statistical tests are run to find significant effects. The applicable tests used are determined by experimental designs and the form and distribution of the data. Generally, ANOVA tests and T-test are usually used when comparing distributions between different conditions. Structural Equation Models (SEM) or multi-level models are used for mediation analysis. More details of statistic tools can be found in Section IV-C. Distributional assumption checks should be applied. When Likert-type data is collected as in most of the questionnaires, non-parametric tests such as paired Wilcoxon signed-rank test, or Kruskal-Wallis H test for multiple groups can be used to avoid normality assumptions.

If multiple measures are aggregated into a single instrument, it is important to assess the validity of this aggregation with

[1] [Online]. Available: https://aspredicted.org

Before the User Study

Measured Quantity
- measurable quantity according to a general definition or application-based measure
- conceptualize measures in social sciences and in the context of XAI

Control Group
- include baseline without explanations
- include baseline with other (e.g., random) explanations in case of a comparative study

Proxy Task
- carefully balanced difficulty
- close to the anticipated real-world application
- participants might have some background knowledge of the task but not too much to bias their judgment or performance

Measurements
- measurements are aligned to definitions
- independent measurements vs. participant-interdependent measurements
- aggregation of the measured quantities across participants
- avoid one-dimensional measures
- pre-register study online to prevent p-hacking
- run a pre-study and expert interviews for initial feedback on the envisioned measures

During the User Study

Experiment Script
- prepare a script with explicit instructions and procedure of the experiment
- obtain consent
- avoid inadvertent cues

Participants
- recruit a representative number of participants for each study subgroup according to the experimental design and required expertise
- adhere to the real-world distribution of users according to the application at hand

Data Quality
- attention checks during a long survey
- preventing order and learning effect (within-subject)

After the User Study

Statistical Verification
- comparison between conditions: use ANOVA, t-test, etc.
- many dependent variables: use SEM, multi-level model, etc.
- distributional assumption checks for different statistical tests (e.g., normality tests)
- Compute measures like Cronbach's α for verifying scales and Fleiß's κ for detecting randomness in the answers of participants or consistent behavior, i.e., agreement beyond chance in their answers

Effect Verification
- correlation between the subjective and behavioral measures
- agreement or disagreement with expected (e.g., theoretical) results

Reporting
- (aggregated) measurements (across participants) and their statistical validity
- model and hyperparameters for reproducibility
- number of participants, characteristics of participants, recruitment
- exact conditions
- analysis methods (name and parameters of the test)

Fig. 3. Summary cards of the guidelines extracted from past XAI user studies.

reliability measures such as the tau-equivalent reliability (also known as Cronbach's α). For example, if objective and subjective measures of a quantity, such as understanding are combined, it is necessary to verify that there is sufficient agreement. If multiple items (e.g., data samples or visualizations) are rated by several subjects, statistics such as Cohan's κ as Fleiß's κ for more than two raters [180] can be used to assess agreement beyond chance between these raters and serve as an indication for the reliability of the ratings.

In the final writing phase, it is essential to report sufficient details that allow readers to estimate the explanatory power of the study. On the level of participants, this should include the total number of participants and how many are assigned to each treatment group, their recruitment, consent and incentivization, and the exact treatment conditions they are subjected to. Furthermore, some descriptive statistics of the collected data can help readers assess the characteristics of the adequacy of the statistical tools used. Regarding the analysis, we found it important to mention how the underlying assumptions of the statistical tests used were checked and to mention the exact variant of the test used (e.g., stating "a two-way ANOVA with the independent variables X and Y" is used instead of just mentioning that ANOVA-test is used).

VII. Future Research Directions

Our survey of recent and ongoing XAI research also helps us identify research gaps and distill a few directions for future investigations. In this section, we highlight these directions and summarize our findings.

A. *Towards Increasingly User-Centered XAI*

We advocate that user-centered methods should be used not only to assess XAI solutions (e.g., through user studies) but also to design them (e.g., through user-centered design). By explicitly modeling and involving users in the design phase and not just in a post-hoc manner during the evaluation phase, we expect the development of XAI solutions that better respond to user needs. As discussed in [117], there are two aspects of human-centered AI: (1) AI systems that understand humans with a sociocultural background and (2) AI systems that help humans understand them. The former point can guide the design of AI systems. In this section, we discuss XAI research that leverages this insight.

The process of explaining a machine's decisions to human users can be viewed as a teaching-learning process where the XAI system is the teacher and the human users are the students. From a user-centered perspective, the problem of designing effective teaching methods to enhance the student's (i.e., user's) learning outcomes is essential to human-centered XAI algorithms. To leverage the ability of humans and address unique user's needs, it is important to review studies and findings from psychology and education. These studies provide insights into how humans perceive other intelligent agents (humans or artificial agents) and how they utilize limited information to infer and generalize. Understanding how humans think and learn will help XAI developers build and design systems that are not only informative but also user-friendly to people with different backgrounds. In this section, we discuss three pedagogical frameworks, namely (1) the expectancy-value motivation theory, (2) the theory of mind, and (3) hybrid teaching, to shed light on incorporating such methods in computational approaches. Inspired by existing work in pedagogy and XAI, we provide implications for designing future transparent AI systems and human-centered evaluations.

Expectancy-Value Motivation Theory: Human interaction with XAI interfaces can be viewed as an activity where humans learn about the model's inner workings through explanations and then achieve an understanding of the models. The question of how to enhance the efficiency and the outcome of this human learning process is of high importance [181]. This research problem is widely considered in educational psychology through the lens of expectancy-value motivation theory. For instance, Hulleman et al. [171] propose to utilize *interventions* to increase the perception of usefulness (utility value) to subsequently increase motivation and final performance. Intervention here refers to identifying the relevance of model explanations to the user's own situation, which can be a prompt question while working with the interface. Moreover, when utilizing model explanations in human-AI collaboration, explanations can be seen as a type of "scaffolding" (prompt during a task) proposed in a conceptual framework in education.

Theory of Mind: When interacting with XAI systems, humans form mental models of the machine learning algorithms that reflect their belief of how the algorithms work. The formation of these mental models comes from observing explanations or examples given to the human, who often subconsciously applies the observations in a few examples to the broader understanding of the whole machine learning system. This incredible ability to infer, rationalize, and summarize other intelligent agent's decisions is known as the Theory of Mind (ToM) in psychology. Based on this theory, the Bayesian Theory of Mind (BToM) provides a probabilistic framework to predict inferences that people make about mental states underlying other agents' actions. Recent work, at the intersection of XAI and robotics, indicates that humans also attribute ToM to artificial agents that they observe or interact with. Guided by these user-centered results, several works at the intersection of XAI and robotics have utilized BToM to create a simulated user, and then use it to generate helpful explanations.

Hybrid Teaching: Teaching strategies for the human-to-human setting have been widely studied and many categorizations exist. One way of categorizing these strategies is through the following three concepts: (1) direct teaching, (2) indirect teaching, and (3) hybrid teaching. *Direct teaching* utilizes direct instructions that are teacher-centered, involve clear teaching objectives, and are consistent with classroom organizations. In XAI applications, direct teaching methods generate explanations by selecting representative examples of an agent's decisions to convey the patterns in its policy. In contrast, *indirect teaching* is student-centered and encourages independent learning. In the XAI perspective, methods utilizing indirect teaching provide users with tools to actively and independently explore an AI system. Technically, direct teaching focuses on providing guidance (using a computational approach) to assist users in building an understanding of a machine, whereas indirect teaching (often through a user interface) enables users to address individual learning preferences and mitigate individual confusion about the AI. To leverage the advantages of the two teaching strategies, *hybrid teaching* has been widely used in human-to-human teaching with an emphasis on interactivity. Recent work [182] indicates that hybrid teaching reduces the amount of time for a user to understand an agent's policy compared to direct and indirect teaching, and is more subjectively preferred by the participants. Building on this, future XAI systems can consider using hybrid teaching methods that (i) generate direct instructions to provide guidance to user's understanding of an AI system; and (ii) provide methods to allow users to interact with the agent.

Explanations through Large Language Models (LLMs): The recent rise of Large Language Models [183], [184] naturally opens up new research directions. There is a growing interest in leveraging their unprecedented capabilities [185] to offer explanations for model decisions [186], [187]. Through their natural language interface, LLMs offer the possibility to build interactive explainers [188]. Intriguingly, textual explanations can also be used as subsequent inputs to LLMs which may help to solve subsequent problems and result in superior performance [189]. This technique, referred to as chain-of-thought reasoning [190], opens up an interesting research territory combining interpretability and performance considerations.

B. Open Research Problems

1) Automatic versus Human-Subject Evaluations: With automatic evaluations, we refer to evaluation methods that do not require human subjects, which corresponds to the functionally-grounded metrics discussed in [119], [120]. These metrics aim to test desiderata around the "faithfulness"/"fidelity"/ "truthfulness" of model explanations [120], [121], [191]. Faithfulness of explanations is defined as that explanations are indicative of true important features in the input [191]. The automatic evaluations aim at capturing general objectivity which is independent from downstream tasks, while human evaluations are contextualized with specific use cases. Generally speaking, automatic evaluations and human evaluations tackle different research challenges: the former objectively examines how truly explanations reflect models and the latter one measures how humans perceive models through explanations (although there existing algorithms for automated evaluation designed to align with human evaluations, which we will discuss later). All explanations used in human-subject experiments should have satisfying performance in automatic evaluations, i.e., the explanations should be able to faithfully unbox the model. This verification step is essential to guarantee the validity of the empirical user study and to ensure that users are not tricked by unfaithful explanations. However, in most current human-subject experiments, the functional faithfulness of explanations is not thoroughly verified beforehand. Using unfaithful explanations could lead to the problem that only the placebo effect of explanations is measured. Ideally, a good explanation should be faithful to the model as well as understandable by users.

2) Identifying and Handling Confounders: Existing research underscores the vulnerability of model explanation studies to significant confounding effects. For instance, Papenmeier et al. [155] reveal that user trust can be more influenced by model accuracy than the faithfulness of the explanation itself. Similarly, Yin et al. [192] demonstrate that the accuracy score perceived by users and the one shown to users contribute to trust formation.

A different problem is that good explanations also reveal weaknesses of the model. However, when seeing unexpected explanations, users may express their negative feelings about the model through negative ratings of the explanations. Therefore, good model explanations should help users *calibrate* their trust [26], [193], i.e., trust the model's decision when it is correct but distrust it otherwise. There is a disagreement on how to handle such cases: When evaluating model fairness, several works [2], [27], [38], [73], [75] reckon the increase in perceived fairness as positive, while Dodge et al. [74] define the decrease as positive. Other factors, such as the temporal occurrence of model errors (Nourani et al. [56]), and the dimensions of models (Ross et al. [32], Poursabzi et al. [53]), also come into play.

In summary, these confounding elements suggest that users might be led to put more trust in oversimplified, deceptive, or simply unfaithful explanations. To mitigate this, we recommend meticulous analysis, control and reporting of potential confounders, such as explanation faithfulness and model accuracy, across various test conditions. More advanced measures have been suggested as well. For instance, Schoeffer and Kuehl's [79] propose *appropriate fairness perceptions*, which measures whether people increase or decrease their fairness perceptions depending on the algorithmic fairness of the underlying model. Nevertheless, the thorough investigation of confounding factors remains a challenge. Calibrated measures that are less prone to confounding can be a valuable step forward.

3) Mitigating Personal Biases for XAI: Most XAI techniques and corresponding designed user studies provide *one-size-fits-all* solutions. Individual bias, rooted in a user's mental framework, influences the user's perception of a model. It should be considered in XAI design, development, and evaluation procedures. Several studies that aim to explain reinforcement learning policies utilize cognitive science theories to create a model of the human user [181], [182], [194], [195]. They then generate explanations based on this human model and verify the benefits of tailoring explanations for individual user models. Within the scope of XAI, [196], [197] utilize a Bayesian Teaching framework to capture human perception of model explanations. In user studies, depending on cultural and educational background, participants may likely give different feedback [31]. This kind of personal bias can be mitigated by deploying a large sample size and recruiting participants who are representative of the target audience. We advocate that personal biases should be taken into account in the realm of XAI development.

4) Human-in-the-Loop and Sequential Explanations: In several relevant cases, such as online recommendation systems, users are not only confronted with an explanation once but instead view decisions and potential explanations repeatedly. Recent work in this domain [35] has shown that the order of decisions and explanations may indeed have an effect on user perception and understanding. The AI model may continue to shape the user's mental model over time. The differences between the single-use and the sequential setting still remain to be thoroughly investigated.

5) Proxy Tasks Should Be Close to Real-World Tasks: When using proxy tasks to evaluate models, for instance, to measure subjective understanding, there is a great choice of tasks present in the literature. A good proxy task should have the following features: (1) it has close real-world connections [119]; (2) users or participants have some background knowledge of the task but not too much to affect their judgment or performance during the task; (3) the task is not too complicated to implement or there exists an existing implementation but was used for different purposes (i.e., not used for XAI); and (4) it has connections to existing work. Yet, the link between evaluations through different proxy tasks and real-world applications has not been made very explicit to date. Buçinca et al. [13] show that the outcomes of proxy evaluations can be different from a real-world task. More specifically, the widely accepted proxy tasks, where users are asked to build the mental models of the AI, may not predict the performance in actual decision-making tasks, where users make use of the explanations to assist in making decisions. The results show that users trust different explanations in the proxy task and the actual decision-making task. Therefore, we argue that further research is required to uncover the links between current proxy tasks and on-task performance or to devise new proxy tasks with a verified connection to actual tasks.

6) Simulated Evaluation as a Cost-Efficient Solution: As human-subject experiments are costly to conduct, Chen et al. [198] propose a simulated evaluation framework (`SimEvals`) to select potential explanations for user studies by measuring the predictive information provided by explanations. Concretely, the authors consider three use cases where model explanations are deployed: forward simulation, counterfactual reasoning, and data debugging. Human performance is measured for these three tasks with different explanations. If there is a significant gap in settings of using two types of explanations, the simulated evaluation can also observe such a gap under the same task settings as well. Meanwhile, first attempts to simulate human textual responses in a given context using large language models show that models can provide surprisingly anthropomorphic answers [199]. Undoubtedly and also affirmed by Chen et al. [198], it is not yet realistic to replace human evaluation with the simulated framework as other factors e.g., cognitive biases can affect human decisions. To better simulate human evaluations, more effort should be directed towards modeling human cognitive processes. Concurrently and with appropriate caveats, XAI researchers should also leverage existing and approximate models of human cognition to enable rapid prototyping and assessment of explanations. Section VII-A discusses several candidate human cognition models and highlights recent XAI works [181], [182] that utilize this "Oz-of-Wizard" paradigm.

VIII. Conclusion

In recent years, there has been a proliferation of XAI research in both academia and industry. Explainability is a human-centric property [141] and therefore XAI should be preferably studied by taking humans' feedback into account. In this work, we investigated recent user studies for XAI techniques through a principled literature review. Based on our review, we found out that the effectiveness of XAI in users' interaction with ML models was not consistent across different applications, thus suggesting that there is a strong need for more transparent and comparable human-based evaluations in XAI. Furthermore, relevant disciplines, such as cognitive psychology and social sciences in general, should become an integral part of XAI research.

We comprehensively analyzed the design patterns and findings from previous works. Based on best-practice approaches and measured quantities, we propose a general guideline for human-centered user studies and several future research directions for XAI researchers and practitioners. Thereby, this work represents a starting point for more transparent and human-centered XAI research.

References

[1] C. Panigutti, A. Beretta, F. Giannotti, and D. Pedreschi, “Understanding the impact of explanations on advice-taking: A user study for AI-based clinical decision support systems,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2022, pp. 1–9.

[2] A. I. Anik and A. Bunt, “Data-centric explanations: Explaining training data of machine learning systems to promote transparency,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–13.

[3] M. Colley, B. Eder, J. O. Rixen, and E. Rukzio, “Effects of semantic segmentation visualization on trust, situation awareness, and cognitive load in highly automated vehicles,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–1.

[4] U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, and J. D. Weisz, “Expanding explainability: Towards social transparency in ai systems,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–19.

[5] M. Liao and S. S. Sundar, “How should AI systems talk to users when collecting their personal information? effects of role framing and self-referencing on Human-AI interaction,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–14.

[6] C.-H. Tsai, Y. You, X. Gui, Y. Kou, and J. M. Carroll, “Exploring and promoting diagnostic transparency and explainability in online symptom checkers,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–17.

[7] L. Guo, E. M. Daly, O. Alkan, M. Mattetti, O. Cornec, and B. Knijnenburg, “Building trust in interactive machine learning via user contributed interpretable rules,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2022, pp. 537–548.

[8] J. Ooge, S. Kato, and K. Verbert, “Explaining recommendations in E-learning: Effects on adolescents’ trust,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2022, pp. 93–105.

[9] H. Suresh, K. M. Lewis, J. Guttag, and A. Satyanarayan, “Intuitively assessing ML model reliability through example-based explanations and editing model inputs,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2022, pp. 767–781.

[10] R. Paleja, M. Ghuy, N. Ranawaka Arachchige, R. Jensen, and M. Gombolay, “The utility of explainable AI in ad hoc human-machine teaming,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 610–623.

[11] J. Schaffer, J. O’Donovan, J. Michaelis, A. Raglin, and T. Höllerer, “I can do better than your AI: Expertise and explanations,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2019, pp. 240–251.

[12] X. Wang and M. Yin, “Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2021, pp. 318–328.

[13] Z. Buçinca, P. Lin, K. Z. Gajos, and E. L. Glassman, “Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2020, pp. 454–464.

[14] X. Peng, M. Riedl, and P. Ammanabrolu, “Inherently explainable reinforcement learning in natural language,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 16178–16190.

[15] Y. Zhang, Q. V. Liao, and R. K. Bellamy, “Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making,” in *Proc. Conf. Fairness Accountability Transparency*, 2020, pp. 295–305.

[16] V. Dominguez, P. Messina, I. Donoso-Guzmán, and D. Parra, “The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2019, pp. 408–446.

[17] C. J. Cai, J. Jongejan, and J. Holbrook, “The effects of example-based explanations in a machine learning interface,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2019, pp. 258–262.

[18] M. Millecamp, N. N. Htun, C. Conati, and K. Verbert, “To explain or not to explain: The effects of personal characteristics when explaining music recommendations,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2019, pp. 397–407.

[19] C.-H. Tsai and P. Brusilovsky, “Beyond the ranked list: User-driven exploration and diversification of social recommendation,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2018, pp. 239–250.

[20] T. Li, G. Convertino, R. K. Tayi, and S. Kazerooni, “What data should I protect? recommender and planning support for data security analysts,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2019, pp. 286–297.

[21] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan, “Interpreting interpretability: Understanding data scientists’ use of interpretability tools for machine learning,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–14.

[22] H.-F. Cheng et al., “Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–12.

[23] J. Kunkel, T. Donkers, L. Michael, C.-M. Barbu, and J. Ziegler, “Let me explain: Impact of personal and impersonal explanations on trust in recommender systems,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–12.

[24] D. H. Kim, E. Hoque, and M. Agrawala, “Answering questions about charts and generating visual explanations,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–13.

[25] V. Lai and C. Tan, “On human predictions with explanations and predictions of machine learning models: A case study on deception detection,” in *Proc. ACM Conf. Fairness Accountability Transparency*, 2019, pp. 1–13.

[26] Y. Rong, N. Castner, E. Bozkir, and E. Kasneci, “User trust on an explainable ai-based medical diagnosis support system,” 2022, *arXiv:2204.12230*.

[27] J. Schoeffer, N. Kuehl, and Y. Machowski, ““there is not enough information”: On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making,” 2022, *arXiv:2205.05758*.

[28] U. Ehsan, P. Tambwekar, L. Chan, B. Harrison, and M. O. Riedl, “Automated rationale generation: A technique for explainable AI and its effects on human perceptions,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2019, pp. 263–274.

[29] A. Smith-Renner et al., “No explainability without accountability: An empirical study of explanations and feedback in interactive ML,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–13.

[30] A. Smith-Renner, V. Kumar, J. Boyd-Graber, K. Seppi, and L. Findlater, “Digging into user control: Perceptions of adherence and instability in transparent models,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2020, pp. 519–530.

[31] A. Springer and S. Whittaker, “Progressive disclosure: Empirically motivated approaches to designing effective transparency,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2019, pp. 107–120.

[32] A. Ross, N. Chen, E. Z. Hang, E. L. Glassman, and F. Doshi-Velez, “Evaluating the interpretability of generative models by interactive reconstruction,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–15.

[33] M. Radensky, D. Downey, K. Lo, Z. Popovic, and D. S. Weld, “Exploring the role of local and global explanations in recommender systems,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2022, pp. 1–7.

[34] S. Hadash, M. C. Willemsen, C. Snijders, and W. A. IJsselsteijn, “Improving understandability of feature contributions in model-agnostic explainable AI tools,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2022, pp. 1–9.

[35] M. Chromik, M. Eiband, F. Buchner, A. Krüger, and A. Butz, “I think I get your point, AI! the illusion of explanatory depth in explainable AI,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2021, pp. 307–317.

[36] J. Rebanal, J. Combitsis, Y. Tang, and X. Chen, “XAlgo: A design probe of explaining algorithms’ internal states via question-answering,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2021, pp. 329–339.

[37] U. Kuhl, A. Artelt, and B. Hammer, “Keep your friends close and your counterfactuals closer: Improved learning from closest rather than plausible counterfactual explanations in an abstract setting,” 2022, *arXiv:2205.05515*.

[38] E. Rader, K. Cotter, and J. Cho, “Explanations as mechanisms for supporting algorithmic transparency,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–13.

[39] A. Bell, I. Solano-Kamaiko, O. Nov, and J. Stoyanovich, “It’s just not that simple: An empirical study of the accuracy-explainability trade-off in machine learning for public policy,” in *Proc. ACM Conf. Fairness Accountability Transparency*, 2022, pp. 248–266.

[40] P. Hase and M. Bansal, “Evaluating explainable AI: Which algorithmic explanations help users predict model behavior?,” in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 5540–5552.

[41] H. Schuff, A. Jacovi, H. Adel, Y. Goldberg, and N. T. Vu, “Human interpretation of saliency-based explanation over text,” 2022, *arXiv:2201.11569*, .

[42] S. Bang, P. Xie, H. Lee, W. Wu, and E. Xing, “Explaining a black-box by using a deep variational information bottleneck approach,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 11396–11404.

[43] S. S. Kim, N. Meister, V. V. Ramaswamy, R. Fong, and O. Russakovsky, “HIVE: Evaluating the human interpretability of visual explanations,” in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 280–298.

[44] M. Szymanski, M. Millecamp, and K. Verbert, “Visual, textual or hybrid: The effect of user expertise on different explanations,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2021, pp. 109–119.

[45] G. Plumb, M. Al-Shedivat, Á. A. Cabrera, A. Perer, E. Xing, and A. Talwalkar, “Regularizing black-box models for improved interpretability,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 10526–10536.

[46] W. Zhang and B. Y. Lim, “Towards relatable explainable ai with the perceptual process,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2022, pp. 1–24.

[47] C. Bove, J. Aigrain, M.-J. Lesot, C. Tijus, and M. Detyniecki, “Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2022, pp. 807–819.

[48] A. Abdul, C. von der Weth, M. Kankanhalli, and B. Y. Lim, “COGAM: Measuring and moderating cognitive load in machine learning model explanations,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–14.

[49] K. Natesan Ramamurthy, B. Vinzamuri, Y. Zhang, and A. Dhurandhar, “Model agnostic multilevel explanations,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 5968–5979.

[50] S. Arora, D. Pruthi, N. Sadeh, W. W. Cohen, Z. C. Lipton, and G. Neubig, “Explain, edit, and understand: Rethinking user study design for evaluating model explanations,” in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 5277–5285.

[51] J. Antoran, U. Bhatt, T. Adel, A. Weller, and J. M. Hern ández-Lobato, “Getting a {clue}: A method for explaining uncertainty estimates,” in *Proc. Int. Conf. Learn. Representations*, 2021.

[52] J. Borowski et al., “Exemplary natural images explain {CNN} activations better than state-of-the-art feature visualization,” in *Proc. Int. Conf. Learn. Representations*, 2021.

[53] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, and H. Wallach, “Manipulating and measuring model interpretability,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–52.

[54] A. Alqaraawi, M. Schuessler, P. Weiß, E. Costanza, and N. Berthouze, “Evaluating saliency map explanations for convolutional neural networks: A user study,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2020, pp. 275–285.

[55] M. T. Ribeiro, S. Singh, and C. Guestrin, “Anchors: High-precision model-agnostic explanations,” in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1527–1535.

[56] M. Nourani et al., “Anchoring bias affects mental model formation and user reliance in explainable ai systems,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2021, pp. 340–350.

[57] L. Sixt, M. Schuessler, O.-I. Popescu, P. Weiß, and T. Landgraf, “Do users benefit from interpretable vision? a user study, baseline, and dataset,” in *Proc. Int. Conf. Learn. Representations*, 2022.

[58] A. Chandrasekaran, V. Prabhu, D. Yadav, P. Chattopadhyay, and D. Parikh, “Do explanations make VQA models more predictable to a human?,” in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2018, pp. 1036–1042.

[59] J. Colin, T. Fel, R. Cadene, and T. Serre, “What I cannot predict, I do not understand: A human-centered evaluation framework for explainability methods,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 2832–2845.

[60] H. Shen and T.-H. Huang, “How useful are the machine-generated interpretations to general users? a human evaluation on guessing the incorrectly predicted labels,” in *Proc. AAAI Conf. Hum. Comput. Crowdsourcing*, 2020, pp. 168–172.

[61] C.-K. Yeh, B. Kim, S. O. Arik, C.-L. Li, T. Pfister, and P. Ravikumar, “On completeness-aware concept-based explanations in deep neural networks,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 20554–20565.

[62] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, “Towards automatic concept-based explanations,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 9277–9286.

[63] T. Leemann, Y. Rong, S. Kraft, E. Kasneci, and G. Kasneci, “Coherence evaluation of visual concepts with objects and language,” in *Proc. Int. Conf. Learn. Representations WS*, 2022.

[64] I. Laina, R. Fong, and A. Vedaldi, “Quantifying learnability and describability of visual concepts emerging in representation learning,” *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 13112–13126.

[65] Y. Wang, P. Venkatesh, and B. Y. Lim, “Interpretable directed diversity: Leveraging model explanations for iterative crowd ideation,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2022, pp. 1–28.

[66] D. L. Arendt, N. Nur, Z. Huang, G. Fair, and W. Dou, “Parallel embeddings: A visualization technique for contrasting learned representations,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2020, pp. 259–274.

[67] W. Zhang, M. Dimiccoli, and B. Y. Lim, “Debiased-CAM to mitigate image perturbations with faithful visual explanations of machine learning,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2022, pp. 1–32.

[68] J. Gao, X. Wang, Y. Wang, and X. Xie, “Explainable recommendation through attentive multi-view learning,” in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 3622–3629.

[69] P. Kouki, J. Schaffer, J. Pujara, J. O’Donovan, and L. Getoor, “Personalized explanations for hybrid recommender systems,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2019, pp. 379–390.

[70] C.-H. Tsai and P. Brusilovsky, “Explaining recommendations in an interactive hybrid social recommender,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2019, pp. 391–396.

[71] A. Balayn, N. Rikalo, C. Lofi, J. Yang, and A. Bozzon, “How can explainability methods be used to support bug identification in computer vision models?,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2022, pp. 1–16.

[72] K. Rawal and H. Lakkaraju, “Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 12187–12198.

[73] N. Grgić-Hlača, E. M. Redmiles, K. P. Gummadi, and A. Weller, “Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction,” in *Proc. Wide Web Conf.*, 2018, pp. 903–912.

[74] J. Dodge, Q. V. Liao, Y. Zhang, R. K. Bellamy, and C. Dugan, “Explaining models: An empirical study of how explanations impact fairness judgment,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2019, pp. 275–285.

[75] G. Harrison, J. Hanson, C. Jacinto, J. Ramirez, and B. Ur, “An empirical study on the perceived fairness of realistic, imperfect machine learning models,” in *Proc. Conf. Fairness Accountability Transparency*, 2020, pp. 392–402.

[76] C. Wang et al., “Do humans prefer debiased AI algorithms? a case study in career recommendation,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2022, pp. 134–147.

[77] N. N. Htun, E. Lecluse, and K. Verbert, “Perception of fairness in group music recommender systems,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2021, pp. 302–306.

[78] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt, “‘it’s reducing a human being to a percentage’ perceptions of justice in algorithmic decisions,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–14.

[79] J. Schoeffer and N. Kuehl, “Appropriate fairness perceptions? on the effectiveness of explanations in enabling people to assess the fairness of automated decision systems,” in *Proc. Companion: Companion Pub. Conf. Comput. Supported Cooperative Work Social Comput.*, 2021, pp. 153–157.

[80] T. Donkers, T. Kleemann, and J. Ziegler, “Explaining recommendations by means of aspect-based transparent memories,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2020, pp. 166–176.

[81] F. Hohman, A. Head, R. Caruana, R. DeLine, and S. M. Drucker, “Gamut: A design probe to understand how data scientists understand machine learning models,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–13.

[82] U. Kuhl, A. Artelt, and B. Hammer, “Let’s go to the alien zoo: Introducing an experimental framework to study usability of counterfactual explanations for machine learning,” 2022, *arXiv:2205.03398*.

[83] T. Schneider, J. Hois, A. Rosenstein, S. Ghellal, D. Theofanou-Fülbier, and A. R. Gerlicher, “ExplAIn yourself! transparency for positive UX in autonomous driving,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–12.

[84] S. Choi, K. Aizawa, and N. Sebe, “FontMatcher: Font image paring for harmonious digital graphic design,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2018, pp. 37–41.

[85] P. Le Bras, D. A. Robb, T. S. Methven, S. Padilla, and M. J. Chantler, “Improving user confidence in concept maps: Exploring data driven explanations,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–13.

[86] R. Shang, K. K. Feng, and C. Shah, “Why am I not seeing it? understanding users’ needs for counterfactual explanations in everyday recommendations,” in *Proc. ACM Conf. Fairness Accountability Transparency*, 2022, pp. 1330–1340.

[87] J. Dodge, A. A. Anderson, M. Olson, R. Dikkala, and M. Burnett, “How do people rank multiple mutant agents?,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2022, pp. 191–211.

[88] D. Das and S. Chernova, “Leveraging rationales to improve human task performance,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2020, pp. 510–518.

[89] G. Bansal et al., “Does the whole exceed its parts? the effect of ai explanations on complementary team performance,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–16.

[90] V. Lai, H. Liu, and C. Tan, ““why is’ Chicago’deceptive?,” towards building model-driven tutorials for humans,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–13.

[91] S. Feng and J. Boyd-Graber, “What can ai do for me? evaluating machine learning interpretations in cooperative play,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2019, pp. 229–239.

[92] Y. Alufaisan, L. R. Marusich, J. Z. Bakdash, Y. Zhou, and M. Kantarcioglu, “Does explainable artificial intelligence improve human decision-making?,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 6618–6626.

[93] K. Z. Gajos and L. Mamykina, “Do people engage cognitively with AI? impact of AI assistance on incidental learning,” in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2022, pp. 794–806.

[94] M. Liao, S. S. Sundar, and J. B. Walther, “User trust in recommendation systems: A comparison of content-based, collaborative and demographic filtering,” in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2022, pp. 1–14.

[95] G. Nguyen, D. Kim, and A. Nguyen, “The effectiveness of feature attribution methods and its correlation with automatic evaluation scores,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 26422–26436.

[96] M. R. Taesiri, G. Nguyen, and A. Nguyen, “Visual correspondence-based explanations improve AI robustness and human-AI team accuracy,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 34287–34301.

[97] J. Wei, J. He, K. Chen, Y. Zhou, and Z. Tang, “Collaborative filtering and deep learning based recommendation system for cold start items,” *Expert Syst. Appl.*, vol. 69, pp. 29–39, 2017.

[98] S. Yang, M. Korayem, K. AlJadda, T. Grainger, and S. Natarajan, “Combining content-based and collaborative filtering for job recommendation system: A cost-sensitive statistical relational learning approach,” *Knowl.-Based Syst.*, vol. 136, pp. 37–45, 2017.

[99] Y. Zhang, X. Chen, Q. Ai, L. Yang, and W. B. Croft, “Towards conversational search and recommendation: System ask, user respond,” in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2018, pp. 177–186.

[100] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, “A survey of deep learning techniques for autonomous driving,” *J. Field Robot.*, vol. 37, pp. 362–386, 2020.

[101] H. Cui et al., “Multimodal trajectory predictions for autonomous driving using deep convolutional networks,” in *Proc. Int. Conf. Robot. Automat.*, 2019, pp. 2090–2096.

[102] Y. Rong, C. Han, C. Hellert, A. Loyal, and E. Kasneci, “Artificial intelligence methods in in-cabin use cases: A survey,” *IEEE Intell. Transp. Syst. Mag.*, vol. 14, no. 3, pp. 132–145, May/Jun. 2021.

[103] R. R. Murphy, “Introduction to AI robotics,” *Ind. Robot: An Int. J.*, vol. 28, no. 3, pp. 266–267, 2001.

[104] K. Rajan and A. Saffiotti, “Towards a science of integrated AI and robotics,” *Artif. Intell.*, vol. 247, pp. 1–9, 2017.

[105] S. Wachter, B. Mittelstadt, and L. Floridi, “Transparent, explainable, and accountable AI for robotics,” *Sci. Robot.*, vol. 2, 2017, Art. no. eaan6080.

[106] S. H. Park and K. Han, “Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction,” *Radiology*, vol. 286, pp. 800–809, 2018.

[107] J. A. Sidey-Gibbons and C. J. Sidey-Gibbons, “Machine learning in medicine: A practical introduction,” *BMC Med. Res. Methodol.*, vol. 19, 2019, Art. no. 64.

[108] R. Vaishya, M. Javaid, I. H. Khan, and A. Haleem, “Artificial intelligence (AI) applications for COVID-19 pandemic,” *Diabetes Metabolic Syndrome: Clin. Res. Rev.*, vol. 14, pp. 337–339, 2020.

[109] X. Dastile, T. Celik, and M. Potsane, “Statistical and machine learning models in credit scoring: A systematic literature survey,” *Appl. Soft Comput.*, vol. 91, 2020, Art. no. 106263.

[110] M. Ala’raj, M. F. Abbod, M. Majdalawieh, and L. Jum’a, “A deep learning model for behavioural credit scoring in banks,” *Neural Comput. Appl.*, vol. 34, pp. 5839–5866, 2022.

[111] P. M. Addo, D. Guegan, and B. Hassani, “Credit risk analysis using machine and deep learning models,” *Risks*, vol. 6, no. 2, p. 38, 2018.

[112] N. Van Berkel, J. Goncalves, D. Hettiachchi, S. Wijenayake, R. M. Kelly, and V. Kostakos, “Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study,” in *Proc. ACM Hum.-Comput. Interact.*, vol. 3, pp. 1–21, 2019.

[113] T. Sourdin, “Judge V robot?: Artificial intelligence and judicial decision-making,” *Univ. New South Wales Law J.*, vol. 41, no. 4, pp. 1114–1133, 2018.

[114] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, “Mitigating bias in algorithmic hiring: Evaluating claims and practices,” in *Proc. Conf. Fairness Accountability Transparency*, 2020, pp. 469–481.

[115] P. Tambe, P. Cappelli, and V. Yakubovich, “Artificial intelligence in human resources management: Challenges and a path forward,” *California Manage. Rev.*, vol. 61, pp. 15–42, 2019.

[116] D. Castelvecchi, “Can we open the black box of AI?,” *Nature News*, vol. 538, pp. 20–23, 2016.

[117] M. O. Riedl, “Human-centered artificial intelligence and machine learning,” *Hum. Behav. Emerg. Technol.*, vol. 1, pp. 33–36, 2019.

[118] U. Ehsan and M. O. Riedl, “Human-centered explainable AI: Towards a reflective sociotechnical approach,” in *Proc. Int. Conf. Human-Comput. Interact.*, 2020, pp. 449–466.

[119] F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” 2017, *arXiv: 1702.08608*.

[120] M. Nauta et al., “From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI,” *ACM Comput. Surv.*, vol. 55, pp. 1–42, 2023.

[121] R. Tomsett, D. Harborne, S. Chakraborty, P. Gurram, and A. Preece, “Sanity checks for saliency metrics,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 6021–6029.

[122] Y. Rong, T. Leemann, V. Borisov, G. Kasneci, and E. Kasneci, “A consistent and efficient evaluation strategy for attribution methods,” in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 18770–18795.

[123] D. Nguyen, “Comparing automatic and human evaluation of local explanations for text classification,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2018, pp. 1069–1078.

[124] G. Hoffman, “Evaluating fluency in human–robot collaboration,” *IEEE Trans. Human-Mach. Syst.*, vol. 49, no. 3, pp. 209–218, Jun. 2019.

[125] Workshop, “ExSS-ATEC: Explainable smart systems for algorithmic transparency in emerging technologies,” in *Proc. 25th Int. Conf. Intell. User Interfaces Companion*, vol. 1, 2020.

[126] S. Mohseni, N. Zarei, and E. D. Ragan, “A multidisciplinary survey and framework for design and evaluation of explainable AI systems,” *ACM Trans. Interact. Intell. Syst. (TiiS)*, vol. 11, no. 3/4, pp. 1–45, 2021.

[127] Q. Yang, N. Banovic, and J. Zimmerman, “Mapping machine learning advances from HCI research to reveal starting places for design innovation,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–11.

[128] A. Adadi and M. Berrada, “Peeking inside the black-box: A survey on explainable artificial intelligence (XAI),” *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[129] A. B. Arrieta et al., “Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Inf. Fusion*, 2020, vol. 58, pp. 82–115.

[130] W. Samek and K.-R. Müller, “Towards explainable artificial intelligence,” in *Proc. Explainable AI: Interpreting Explaining Visualizing Deep Learn.*, 2019, pp. 5–22.

[131] N. Burkart and M. F. Huber, “A survey on the explainability of supervised machine learning,” *J. Artif. Intell. Res.*, vol. 70, pp. 245–317, 2021.

[132] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, “Machine learning interpretability: A survey on methods and metrics,” *Electronics*, vol. 8, 2019, Art. no. 832.

[133] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, “Explaining explanations: An overview of interpretability of machine learning,” in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Analytics*, 2018, pp. 80–89.

[134] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, “Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–28.

[135] G. Montavon, W. Samek, and K.-R. Müller, “Methods for interpreting and understanding deep neural networks,” *Digit. Signal Process.*, vol. 73, pp. 1–15, 2018.

[136] A. Das and P. Rad, “Opportunities and challenges in explainable artificial intelligence (XAI): A survey,” 2020, *arXiv: 2006.11371*.

[137] G. Joshi, R. Walambe, and K. Kotecha, “A review on explainability in multimodal deep neural nets,” *IEEE Access*, vol. 9, pp. 59800–59821, 2021.

[138] R. Moraffah, M. Karami, R. Guo, A. Raglin, and H. Liu, “Causal interpretability for machine learning-problems, methods and evaluation,” *ACM SIGKDD Explorations Newslett.*, vol. 22, pp. 18–33, 2020.

[139] I. Nunes and D. Jannach, “A systematic review and taxonomy of explanations in decision support and recommender systems,” *User Model. User-Adapted Interact.*, vol. 27, pp. 393–444, 2017.

[140] Z. C. Lipton, “The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery,” *Queue*, vol. 16, pp. 31–57, 2018.

[141] Q. V. Liao and K. R. Varshney, “Human-centered explainable AI (XAI): From algorithms to user experiences,” 2021, *arXiv:2110.10790*.

[142] V. Lai, C. Chen, Q. V. Liao, A. Smith-Renner, and C. Tan, “Towards a science of Human-AI decision making: A survey of empirical studies,” 2021, *arXiv:2112.11471*.

[143] J. J. Ferreira and M. S. Monteiro, “What are people doing about XAI user experience? a survey on ai explainability research and practice,” in *Proc. Int. Conf. Hum.-Comput. Interact.*, 2020, pp. 56–73.

[144] N. Bevan, “International standards for HCI and usability,” *Int. J. Hum.-Comput. Stud.*, vol. 55, pp. 533–552, 2001.

[145] W. Iso, “9241–11: 1998, Ergonomic requirements for work with visual display terminals (VDTs)-Part 11: Guidance on usability,” *Int. Org. Standardization*, vol. 45, no. 9, 1998.

[146] M. T. Ribeiro, S. Singh, and C. Guestrin, ““Why should I trust you?,” explaining the predictions of any classifier,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1135–1144.

[147] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4768–4777.

[148] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[149] P. Voigt and A. Von dem Bussche, “The EU general data protection regulation (GDPR),” in *A Practical Guide*, 1st ed., Berlin, Germany: Springer, 2017.

[150] B. Goodman and S. Flaxman, “European union regulations on algorithmic decision-making and a “right to explanation”,” *AI Mag.*, vol. 38, no. 3, pp. 50–57, 2017.

[151] C. Molnar, “Interpretable machine learning,” pp. 26–27, 2020.

[152] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nat. Mach. Intell.*, vol. 1, pp. 206–215, 2019.

[153] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, “Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission,” in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2015, pp. 1721–1730.

[154] C. Panigutti, A. Perotti, and D. Pedreschi, “Doctor XAI: An ontology-based approach to black-box sequential data classification explanations,” in *Proc. Conf. Fairness Accountability Transparency*, 2020, pp. 629–639.

[155] A. Papenmeier, G. Englebienne, and C. Seifert, “How model accuracy and explanation fidelity influence user trust,” 2019, *arXiv: 1907.12652*.

[156] J. van der Waa, E. Nieuwburg, A. Cremers, and M. Neerincx, “Evaluating XAI: A comparison of rule-based and example-based explanations,” *Artif. Intell.*, vol. 291, 2021, Art. no. 103404.

[157] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, “Machine learning for medical imaging,” *Radiographics*, vol. 37, no. 2, pp. 505–515, 2017.

[158] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, “Foundations for an empirically determined scale of trust in automated systems,” *Int. J. Cogn. Ergonom.*, vol. 4, pp. 53–71, 2000.

[159] B. Kim et al., “Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV),” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2668–2677.

[160] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell, “Explaining the user experience of recommender systems,” in *User Modeling User-Adapted Interaction*. Berlin, Germany: Springer, 2012.

[161] B. Y. Lim and A. K. Dey, “Assessing demand for intelligibility in context-aware applications,” in *Proc. 11th Int. Conf. Ubiquitous Comput.*, 2009, pp. 195–204.

[162] S. G. Hart and L. E. Staveland, “Development of NASA-TLX (task load index): Results of empirical and theoretical research,” *Adv. Psychol.*, vol. 52, pp. 139–183, 1988.

[163] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, “Metrics for explainable AI: Challenges and prospects,” 2018, *arXiv: 1812.04608*.

[164] A. Holzinger, A. Carrington, and H. Müller, “Measuring the quality of explanations: The system causability scale (SCS),” *KI-Künstliche Intelligenz*, 2020.

[165] A. Gegenfurtner, E. Lehtinen, and R. Säljö, “Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains,” *KI-Ku nstliche Intelligenz*, vol. 34, no. 2, pp. 193–198, 2020.

[166] K. Cotter, J. Cho, and E. Rader, “Explaining the news feed algorithm: An analysis of the “news feed FYI,” blog,” in *Proc. CHI Conf. Extended Abstr. Hum. Factors Comput. Syst.*, 2017, pp. 1553–1560.

[167] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, “Designing theory-driven user-centric explainable AI,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–15.

[168] L. Rozenblit and F. Keil, “The misunderstood limits of folk science: An illusion of explanatory depth,” *Cogn. Sci.*, vol. 26, pp. 521–562, 2002.

[169] G. Hoffman and X. Zhao, “A primer for conducting experiments in human–robot interaction,” *ACM Trans. Human-Robot Interact.*, vol. 10, pp. 1–31, 2020.

[170] J. Eccles, “Expectancies, values and academic behaviors,” *Achievement Achievement Motives*, vol. 58, pp. 58–74, 1983.

[171] C. S. Hulleman, J. J. Kosovich, K. E. Barron, and D. B. Daniel, “Making connections: Replicating and extending the utility value intervention in the classroom,” *J. Educ. Psychol.*, vol. 109, 2017, Art. no. 387.

[172] F. G. Paas, “Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach,” *J. Educ. Psychol.*, vol. 84, pp. 429–434, 1992.

[173] K. Ouwehand, A. V. D. Kroef, J. Wong, and F. Paas, “Measuring cognitive load: Are there more valid alternatives to likert rating scales?,” *Front. Educ., Frontiers Educ.*, vol. 6, p. 702616, 2021.

[174] J. P. Simmons, L. D. Nelson, and U. Simonsohn, “Pre-registration: Why and how,” *J. Consum. Psychol.*, vol. 31, pp. 151–162, 2021.

[175] U. Simonsohn, L. D. Nelson, and J. P. Simmons, “P-curve: A key to the file-drawer,” *J. Exp. Psychol.: Gen.*, vol. 143, pp. 534–547, 2014.

[176] K. A. Ericsson and H. A. Simon, *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA, USA: MIT Press, 1984.

[177] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, San Francisco, CA, USA: Academic, 2013.

[178] S. Dhanorkar, C. T. Wolf, K. Qian, A. Xu, L. Popa, and Y. Li, “Who needs to know what, when?: Broadening the explainable AI (XAI) design space by looking at explanations across the AI lifecycle,” in *Proc. Des. Interactive Syst. Conf.*, 2021, pp. 1591–1602.

[179] F. Y. Kung, N. Kwok, and D. J. Brown, “Are attention check questions a threat to scale validity?,” *Appl. Psychol.*, vol. 67, pp. 264–283, 2018.

[180] J. L. Fleiss, “Measuring nominal scale agreement among many raters,” *Psychol. Bull.*, vol. 76, pp. 378–382, 1971.

[181] I. Lage, D. Lifschitz, F. Doshi-Velez, and O. Amir, “Exploring computational user models for agent policy summarization,” in *IJCAI: Proc. Conf.*, 2019, Art. no. 1401.

[182] P. Qian and V. Unhelkar, “Evaluating the role of interactivity on improving transparency in autonomous agents,” in *Proc. 21st Int. Conf. Auton. Agents Multiagent Syst.*, 2022, pp. 1083–1091.

[183] A. Radford et al., “Language models are unsupervised multitask learners,” *OpenAI Blog*, vol. 1, no. 8, 2019, Art. no. 9.

[184] ChatGPT, Introducing, “OpenAI,” 2023. Accessed: Feb. 17, 2023. [Online]. Available: https://openai.com/blog/chatgpt

[185] S. Bubeck et al., “Sparks of artificial general intelligence: Early experiments with GPT-4,” 2023, *arXiv:2303.12712*.

[186] W. Zhou et al., “Towards interpretable natural language understanding with explanations as latent variables,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 6803–6814.

[187] S. Wiegreffe, J. Hessel, S. Swayamdipta, M. Riedl, and Y. Choi, “Reframing Human-AI collaboration for generating free-text explanations,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2022, pp. 632–658.

[188] S. Wang, Z. Zhao, X. Ouyang, Q. Wang, and D. Shen, “Chatcad: Interactive computer-aided diagnosis on medical image using large language models,” 2023, *arXiv:2302.07257*.

[189] N. F. Rajani, B. McCann, C. Xiong, and R. Socher, “Explain yourself! leveraging language models for commonsense reasoning,” in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 4932–4942.

[190] J. Wei et al., “Chain-of-thought prompting elicits reasoning in large language models,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 24824–24837.

[191] D. Alvarez Melis and T. Jaakkola, “Towards robust interpretability with self-explaining neural networks,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 7786–7795.

[192] M. Yin, J. Wortman Vaughan, and H. Wallach, “Understanding the effect of accuracy on trust in machine learning models,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–12.

[193] A. Bussone, S. Stumpf, and D. O’Sullivan, “The role of explanations on trust and reliance in clinical decision support systems,” in *Proc. Int. Conf. Healthcare Inform.*, 2015, pp. 160–169.

[194] C. Baker, R. Saxe, and J. Tenenbaum, “Bayesian theory of mind: Modeling joint belief-desire attribution,” in *Proc. Annu. Meeting Cogn. Sci. Soc.*, vol. 33, no. 33, 2011.

[195] S. H. Huang, D. Held, P. Abbeel, and A. D. Dragan, “Enabling robots to communicate their objectives,” *Auton. Robots*, vol. 43, pp. 309–326, 2019.

[196] S. C.-H. Yang, N. E. T. Folke, and P. Shafto, “A psychological theory of explainability,” in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 25007–25021.

[197] S. C.-H. Yang, W. K. Vong, R. B. Sojitra, T. Folke, and P. Shafto, “Mitigating belief projection in explainable artificial intelligence via Bayesian teaching,” *Sci. Rep.*, vol. 11, 2021, Art. no. 9863.

[198] V. Chen, N. Johnson, N. Topin, G. Plumb, and A. Talwalkar, “Use-case-grounded simulations for explanation evaluation,” 2022, *arXiv:2206.02256*.

[199] G. Aher, R. I. Arriaga, and A. T. Kalai, “Using large language models to simulate multiple humans,” 2022, *arXiv:2208.10264*.

Yao Rong received the MSc degree in electrical and computer engineering from the Technical University of Munich, Germany, in 2019. She is currently working toward the doctoral degree with the Human-Centered Technologies for Learning Group, the Technical University of Munich. From 2022 to 2023, she served as a visiting scholar with the DATA Lab, Rice University. Her research interests lie in human-centered AI, explainable AI, and human-AI interaction technologies.

Tobias Leemann received the MSc degree from the University of Erlangen-Nuremberg, Germany, in 2020. He is currently working toward the PhD degree with the University of Tübingen, Germany where his research is focused on trustworthy machine learning. Specifically, his research interests include the quality assessment of interpretability techniques and the intersections of interpretability, fairness and privacy.

Thai-Trang Nguyen is graduated with a BSc degree in computer science from the University of Tübingen, Germany. She is currently working toward the MSc degree with the same university. Furthermore, she served as a research assistant, the Human-Computer Interaction group from 2019 to 2022.

Lisa Fiedler is currently working toward the BSc degree in media informatics from the University of Tübingen, Germany. Additionally, she works as a student assistant for the Human-Computer Interaction Group at the University of Tübingen.

Peizhu Qian is currently working toward the PhD degree in computer science with Rice University, USA working with Dr. Vaibhav Unhelkar on problems in human-robot interaction, robot transparency, and explainable AI. Her research interest lies in building a mutual understanding between a robot and its human collaborators. Her work applies psychology theories to computational frameworks, enabling robots to communicate their objectives.

Vaibhav Unhelkar received the MS degree in aeronautics and astronautics and the PhD degree in autonomous systems, in 2015 and 2020, respectively. He is an assistant professor of computer science with Rice University, USA where he leads a research group in the emerging area of Human-Centered AI and Robotics. Unhelkar earned his undergraduate degree in aerospace engineering from the Indian Institute of Technology in Bombay, in 2012. From the Massachusetts Institute of Technology, where he worked in the Computer Science and Artificial Intelligence Laboratory (CSAIL).

Tina Seidel received the diploma degree in psychology from the University of Regensburg (Germany) and Vanderbilt University Nashville (USA), in 1998, and the PhD degree with excellence, in 2002 from the Leibniz Institute for Science and Mathematics Education Kiel (Germany). She holds the Friedl Schoeller Chair for Educational Psychology with the School of Social Sciences and Technology, Technical University of Munich, Germany. Her research focuses on teaching and teacher education. She has established a Teacher Research & Training Simulation Center that conducts several research projects funded by the German Science Foundation and the German Federal Ministry of Education and Research.

Gjergji Kasneci received the MSc degree in computer science and mathematics from the University of Marburg, in 2005, and the PhD degree from the University of Saarland - while with the Max Planck Institute - in 2009. He then worked with Microsoft Research Cambridge, the Hasso Plattner Institute, and SCHUFA Holding AG, where he served as CTO from 2017 to 2022. Between 2018 and 2023, he led the Data Science and Analytics Group with the University of Tübingen as an Honorary professor. In 2023, Gjergji Kasneci was appointed professor of Responsible Data Science with the Technical University of Munich.

Enkelejda Kasneci received the PhD degree in computer science from the University of Tübingen, in 2013. She was postdoctoral researcher and a Margarete-von-Wrangell Fellow with the University of Tübingen. She is a distinguished professor for Human-Centered Technologies for Learning with the Technical University of Munich and Core Member of the Munich Data Science Institute. Her research evolves around Human-Centered Technologies and AI systems that sense and infer the user’s cognitive state, the level of task-related expertise, actions, and intentions based on multimodal data and provide information for media and assistive technologies in many activities of everyday life, and especially in the context of learning.