Towards Human-centered Explainable AI: User Studies for Model Explanations

YAO RONG, University of Tübingen, Germany
TOBIAS LEEMANN, University of Tübingen, Germany
THAI-TRANG NGUYEN, University of Tübingen, Germany
LISA FIEDLER, University of Tübingen, Germany
PEIZHU QIAN, Rice University, USA
VAIBHAV UNHELKAR, Rice University, USA
TINA SEIDEL, Technical University of Munich, Germany
GJERGJI KASNECI, University of Tübingen, Germany
ENKELEJDA KASNECI, Technical University of Munich, Germany

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 HCI and AI researchers conduct user studies in XAI applications based on a systematic literature review. After identifying and thoroughly analyzing 85 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, fairness, usability*, and *human-AI team 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.

1 INTRODUCTION

Artificial Intelligence (AI) is driving digital transformation and already an integral part of various everyday technologies. Recent developments in AI are essential to progress in fields such as recommendation systems [121, 240, 246, 251, 258], autonomous driving [43, 58, 89, 100, 197] or robotics [167, 190, 238]. Moreover, AI's success story has not excluded high-stakes decision making tasks like medical diagnosis [10, 73, 184, 214, 232], credit scoring [4, 6, 61, 177, 249], jurisprudence [13, 67, 222, 233] or recruiting and hiring decisions [30, 189, 226, 259], However, the behavior and decision making processes of modern AI systems are often not understandable, so they are frequently considered black boxes. Deploying these black-box models has proven rather difficult in practice, as a transparent and trustworthy AI system is required by both developers (to better identify and fix bugs) 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 [71, 195]. While a huge number of model explanations are

Authors' addresses: Yao Rong, yao.rong@uni-tuebingen.de, University of Tübingen, Sand 14, Tübingen, Germany, 72076; Tobias Leemann, University of Tübingen, Sand 14, Tübingen, Germany, 72076, tobias.leemann@uni-tuebingen.de; Thai-trang Nguyen, University of Tübingen, Sand 14, Tübingen, Germany, 72076, thai-trang.nguyen@student.uni-tuebingen.de; Lisa Fiedler, University of Tübingen, Sand 14, Tübingen, Germany, 72076, lisa.fiedler@student.uni-tuebingen.de; Peizhu Qian, Rice University, 6100 Main St, Houston, USA, 77005, pq3@rice.edu; Vaibhav Unhelkar, Rice University, 6100 Main St, Houston, USA, 77005, vaibhav.unhelkar@rice.edu; Tina Seidel, Technical University of Munich, Arcisstraße 21, Munich, Germany, 80333, tina.seidel@tum.de; Gjergji Kasneci, University of Tübingen, Sand 14, Tübingen, Germany, 72076, gjergji.kasneci@uni-tuebingen.de; Enkelejda Kasneci, Technical University of Munich, Arcisstraße 21, Munich, Germany, 80333, enkelejda.kasneci@tum.de.

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 [66]. 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 [170] for review), however the difficult comparability between different automatic evaluation measures is a common problem [198, 228]. Another drawback of automated measures is that there is no guarantee that they truly reflect humans' preferences [96, 171]. 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 [64]. However, only a small portion (about 20 %) of XAI evaluation projects consider human subjects [170]. There exist efforts in developing taxonomies or introducing the definitions or implications of different human-centric evaluations [51, 103, 161], but the recent generation of user studies and their findings have not been systematically discussed yet. Moreover, Yang et al. [250] point out that XAI is growing separately and treated differently in different communities (e.g., ML and HCI). Hence, effective guidance in XAI user study design is crucial to better let both XAI algorithm 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.

As user studies in XAI require intersectional knowledge across the disciplines of artificial intelligence/machine learning (AI/ML) and human-computer interaction (HCI) as a prerequisite, the scope of this paper covers the main venues in these fields. We identified a total of 85 core papers for this survey (see Table 1 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 of this paper:

(1) To offer an overview of the foundational work of user studies in XAI, we have investigated references of all 85 core papers in a data-driven manner. Likewise, we have 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 3).
(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 hence inspiration of how to collect human assessment (Section 4).
(3) We discuss the impact of using explanations on different aspects of user experience (Section 5), which can serve as an overview of the effectiveness of the current XAI technology and as a summary of the state-of-the-art.
(4) Based on the examined user study details and their best-practice findings, we synthesize and propose guidelines for designing an effective user study for XAI (Section 6).
(5) Beyond the user study design, we discuss potential paradigms of AI systems understanding humans in the contextual of e.g., theory of minds, as well as other future research directions (Section 7).

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.

Trust		[14, 34, 53, 70, 92, 149, 175, 180, 181, 206, 225, 231, 243] [37, 45, 64, 72, 115, 120, 136, 140, 148, 159, 196, 209, 219, 220, 229]
Fairness		[14, 28, 63, 88, 94, 108, 188, 209, 241]
Understanding	subjective	[26, 34, 37, 45, 50, 64, 72, 92, 93, 96, 130, 187, 188, 191, 199, 211, 243]
	objective	[2, 7, 16, 18, 26, 31, 32, 34, 42, 45, 50, 96, 169, 173, 185, 193, 199, 218, 243, 257]
	explanation model	[84, 115, 142, 147, 169, 244, 252, 257]
Usability	workload	[2, 17, 53, 64, 115]
	helpfulness	[2, 34, 82, 173, 244, 256, 257]
	satisfaction	[32, 64, 92, 128, 159, 181, 219, 229–231]
	debugging	[21, 120, 185, 218]
	ease of use and others	[2, 17, 21, 34, 48, 53, 65, 106, 115, 120, 130, 131, 143, 148, 181, 199, 207, 220, 244]
Human-AI Team		[8, 23, 34, 60, 75, 81, 139, 140, 180]

Table 1. Overview of the core papers containing user studies in XAI grouped by categories of measurements. Based on different measures, one paper can belong to different groups.

2 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 [3, 19, 35, 41, 85, 204], the authors aim to cover many facets of XAI technologies ranging from problem definitions, goals, AI/ML model explanations to evaluation measures, while in [1] 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, [59, 113, 163] investigate explanations for Deep Neural Networks (DNNs), where models often take images as input [59, 163]. Joshi et al. [113], 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. [164] provide a literature review for causal explanations. A systematic literature review on explanations for advice-giving systems is conducted in [174]. Among those works, 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 [66, 154, 170]. There have been several attempts to organize the discourse on the faceted concepts and evaluation metrics for XAI models. For example, Doshi-Velez and Kim [66] start by defining interpretability and argue that it can confirm other important desiderata of ML systems such as fairness, trust, causality, and usability, which should be assessed as the quality of model explanations. Lipton [154] addresses these desiderata as well as the transparency of models. Recently, Nauta et al. [170] refine the evaluated properties resulting in 12 concepts such as compactness and correctness.

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 [66, 170]. Although humans are the ultimate target of explainability, assessing model explanations with human subjects is not an easy task due to high requirements on the experimental design and the extensive consumption of time and other resources [66]. Many researchers seek for solutions to evaluate explanations automatically. A comprehensive literature review of these functionally-grounded evaluation methods (without human subjects) can be found in [170].

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 [66, 151]. For instance, Chromik and Schuessler [51] propose a taxonomy on XAI evaluations involving humans. Mohseni et al. [161] 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 [103] 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 in decision making and to benefit end users. Kenny et al. [116] propose to partition a user's mental model into three sub-models, which are user models of the domain, of the AI system and of the explanation. When observing performance improvement in the user study for an AI system, this may not reflect the advancement in the user's model of the AI, but the improvement comes from the user's model of the domain. Thus, it is necessary to separate the three models into independent variables of user studies. A recent review by Lai et al. [138] considers studies on collaborative Human-AI decision making, which may include AI agents providing explanations. The success in Human-AI decision making tasks can be seen as one amongst many other ways to evaluate the effect of explanations. Contrarily, our work aims to provide a comprehensive overview of useful measures, that is not limited to this category. Ferreira and Monteiro [77] 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.

As suggested by Doshi-Velez and Kim [66], a human-subject experiment needs to be designed sophisticatedly to reduce confounding factors. Moreover, Buçinca et al. [34] validate that a proxy task used in explanation evaluation and a real-world task may yield different conclusions. Therefore, a principled design of user studies for XAI evaluation is essential for fairly and transparently comparing them. In this work, we perform a thorough literature review across disciplines in user studies of XAI applications. Our goal is to offer XAI researchers and practitioners an overview of the research questions considered in user studies and to provide comprehensive information on experimental design choices. In the end, a practical guideline is presented, which can be used as a starting point for future exploration of human-centric XAI applications.

3 FOUNDATION AND IMPACT OF XAI USER STUDIES

Before diving into the details of the selected papers (i.e., core papers), we start by inspecting foundational works on which user studies are grounded (i.e., their references). This reveals research topics or works around human-centered XAI and important definitions for measuring different aspects of model interpretability. Furthermore, we seek to gain a better understanding of the core papers' impact by exploring follow-up papers that contain citations to at least one core paper, which reveals trends in human-centered XAI. Because their references and citations constitute more than eight thousand papers in total, we deploy an automatic approach to extract the research topics of those papers, thus enabling further visualizations. This type of bibliometric analysis is well-established in literature research and has been used in works such as [1].

In this section, we first explain how we collected the core papers and present the implementation details of the automatic analysis (and visualization) in Section 3.1. Subsequently, the foundations (Section 3.2) and impact (Section 3.3) of core papers are thoroughly discussed.

3.1 Method

We decided to collect highly relevant papers dealing with XAI user studies in the scope of several impactful conferences and venues. These conferences are widely acknowledged as highly impactful in both the HCI and AI domains. Specifically, we included the recent five years of CHI, IUI, UIST, CSCW, FA(cc)T, ICML, ICRL, NeurIPS, AAAI. In addition, for completeness reasons, we also ran several search queries involving keywords from the two groups "explainable AI" and "user study", as listed in the Table 2. We first 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 do not fulfill the following two criteria: (1) deploying explainable models or techniques; (2) conducting an assessment with human subjects.

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

Table 2. Keywords for our paper search query. Two groups of keywords were used.

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 [66], *interpretability* in ML is defined as the ability to explain or present model predictions in understandable terms to a human. The authors argue that interpretability is used to confirm other important desiderata of AI models such as trust, fairness, or usability. During a profound study of the relevant literature that was previously selected, we identified five sensible categories, that are derived from the considered dependent variables user studies (desiderata of interpretability). These five categories are **trust, fairness, understanding, usability**, and **human-AI team performance**. In Table 1, the studied papers are categorized according to the measured quantities. Intuitively, trust, fairness and understanding refer to whether users perceive models as trustworthy, fair and understandable through model explanations. In usability, different aspects are measured, for instance, whether the system is easy to use or how much cognitive load it requires. The aspect "debugging" in "usability" relates to the question whether explanations can help users debug models. Human-AI team performance is related to scenarios where the AI system provides its predictions, but humans retain the final decisions [23]. In this case, model explanations are deployed to reach a performance superior to that of the AI system or the human decision-maker alone. We argue that differentiation and categorization along these dimensions are most reasonable to study the large body of literature, because the dependent variable of interest is one of the first design choices and many other aspects such as measures are direct implications of this choice.

To perform a data-driven bibliometric analysis of the references and citations for all papers[1], we first collected common references from each category. As we had to deal with a large number of papers, a keyword representing the research topic was assigned to each paper. In this way, we could group the papers according to their content. Concretely, the references were extracted directly from the studied papers (in pdf format). The follow-up works that cite each core paper were retrieved from the Google Scholar platform using the Python API ("Scholarly" [49]). The same API was used to extract abstracts from Google Scholar for all references and citations. Based on the paper titles and abstracts, we

[1]In this section, the word “references” refers to sources contained in the references of one of the core papers while “citations” refers to follow-up works that reference one of the core papers

deployed a keyword extract model KeyBERT [90, 213] to label the paper with one keyword. We visualized papers in a 2-dimensional semantic space according to their keywords using t-SNE [234].

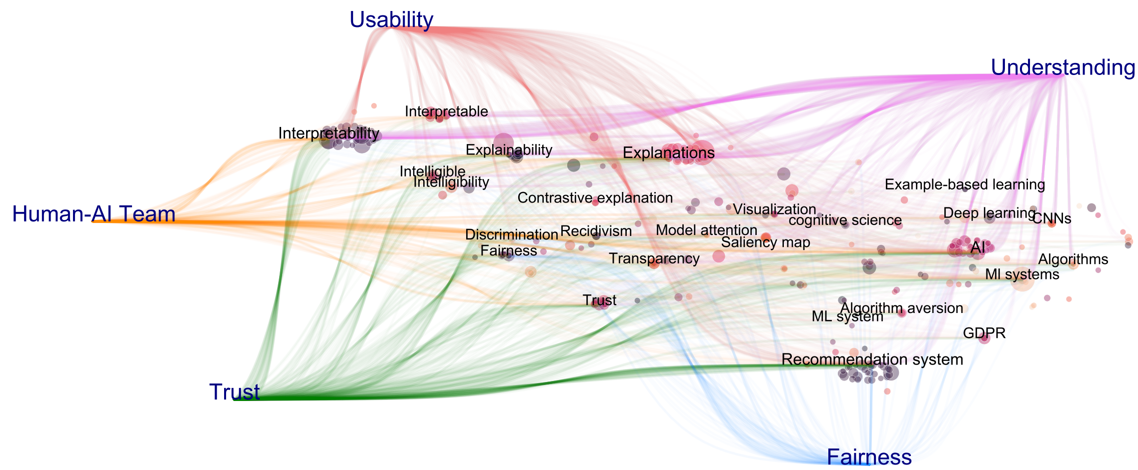

Fig. 1. Illustration of the **foundational** research domains: Each dot represents a referenced paper, whose size reflects the number of studied core papers referring to it. For a clear depiction, only several important research domains are labeled with text. Each line represents a reference link. Five core paper categories are presented in different colors. Lower transparency of lines indicates more links between core papers and their reference papers.

3.2 Foundations of User Studies

We illustrate the research domains that are fundamental to XAI user studies in Figure 1. Note that for presentation clarity, we only visualized works that were used as references in at least three of the core papers. We can see that model explanations and interpretability are essential components, which include papers introducing explanation methods, such as LIME [192] or SHAP values [155]. Papers from all five categories build their work on these methods, indicated by lines with all colors converging at the dots of interpretability. Notably, many research papers appear within the domain of recommender systems, because many XAI user studies are conducted in the context of recommendation solutions. Especially trust and usability of recommender systems are often measured in the corresponding XAI user studies. Besides, trust is an important topic across all five groups of papers. As it is an essential desideratum in user experience, it has been already studied in ML systems going beyond the scope of XAI. Such works still often contain relevant findings. For instance, Yin et al. [253] reveal that showing users the model accuracy can affect perceived trust. The EU's General Data Protection Regulation (GDPR) [237] is mentioned by many core papers due to the ongoing debate on the "right to explanation" [87], which has a huge impact on shaping the modern AI systems towards explainable systems. Examples of particularly common references can be found in Appendix A.

Although the final consumers of model explanations are humans, the well-established research domains involving human understanding, e.g., cognitive science, are underrepresented. Only a small dot labeled "cognitive science" can be observed in Figure 1, representing a paper that proposes to enhance XAI theory with social sciences such as cognitive science and psychology [159]. In combination with the lack of references in the domain of psychology, this indicates that only few XAI user studies attempt to evaluate XAI from human psychological aspects. We highlight a nascent

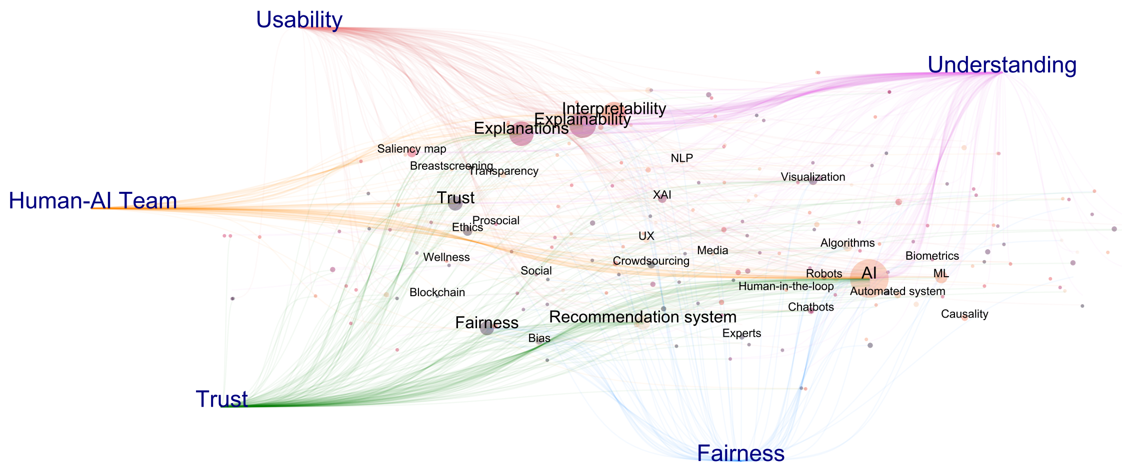

Fig. 2. Illustration of **influenced** research domains: Each dot represents a research topic, whose size refers to the number of papers on the same topic. For a better depiction, only several popular domains were labeled with text. Each line represents a citation link. Five core paper categories are in different colors. Lower transparency of lines indicates more links between core papers and that research topic.

research domain of XAI frameworks based on human cognition and behavior theories [151]. This theoretical guidance can also offer conceptual tools for better evaluating XAI from user perspectives.

3.3 Impact of User Studies

Similarly to foundations in XAI user studies, we are interested in knowing who will eventually benefit from the findings of XAI user studies. Figure 2 demonstrates the "consumers" of the human-centered XAI core papers (i.e., research domains interested influenced by the core papers), with each dot representing a research topic. The size of the dots is determined by the number of citations in the set of core papers obtained from this research area. We observe that user studies on Human-AI Team performance and trust measures can particularly promote the usage and further development of AI systems. Model explainability techniques profit from the user studies' findings as well, because explanations can be improved according to users' needs. The domain of recommendation systems is non-negligible in both figures (for foundations and impact), suggesting that XAI is an inevitable component in modern recommendation systems. Additionally, we notice that human-centered XAI benefits a wide range of application domains, such as media [157], human-in-the-loop AI solutions [247], wellness [76] and medical solutions [39, 227], chatbots [117] and robots [15], prosocial [203] and ethics-oriented solutions [68, 221].

By studying these two aspects (i.e., foundations and impact), we grasp a clear overview of relevant topics in the research landscape of XAI user studies. More importantly, we can better spot the nascent but pertinent areas for future work such as cognition-driven analysis tools in XAI. We release raw data and code for analyses at https://github.com/yaorong0921/hxai-survey.

4 USER STUDY DETAIL 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 4.1), followed by a discussion of application domains and measures with respect to the five measured quantities. The experimental designs, as well as analysis tools, are presented in Section 4.3.

4.1 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 3. 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 [162]. First, there are intrinsically interpretable models, also known as *white-box models*. White-box models can be directly explained by visualizing their parameters. For instance, linear and logistic regression models (cf. [26, 45, 185, 243]), Generalized Additive Models (GAM, cf. [2, 40, 115]), and decision trees (cf. [130, 131]) fall into this category. These types of models are usually applied to tabular data, but exceptionally include other domains such as text data with the Bag-of-Words model in [18], or with a Chatbot in [231]. Second, there are *black-box models* providing no parameter access or are too complex to be explained in a human-understandable way [201]. These include ensembling techniques such as Random Forests (cf. [26, 50, 169, 209]), Gradient-boosted trees (cf. [26, 115]), SVMs (cf. [139, 140]), and neural architectures designed specifically for tabular (cf. [96]), image (cf. [7, 42, 64]), video (cf. [173]), textual (cf. [18, 22, 72, 96]), or audio data (cf. [257]).

As for explanation techniques, we identified five key types in the scope of the surveyed papers (rows of Table 3). Most frequently used are feature-based (attribution) explanations. For instance, [26, 32, 50, 115] deploy SHAP (Shapley additive explanations [155]) and [22, 26, 96, 115, 169, 193] use LIME [192], which are model-agnostic, post-hoc feature attribution methods. Gradient-based attributions for neural networks are used as model-specific explanations (cf. [7, 18, 22, 42]). 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 explanation. Moreover, global and local explanations are also studied in the domain of recommendation systems [187]. Example-based explanations are used for different ML models on a textual classification task (cf. [140]) or deep neural networks in image recognition (cf. [31, 34, 37, 218]). Another particularly popular type of explanations are generated samples in the form of counterfactual explanations, which aim at providing actionable suggestions for attaining a user-preferred prediction through changing certain input features [18, 115, 130, 243]. Concept-based explanations, which use meaningful high-level concepts such as objects or shapes to explain a prediction, are frequent subject of research in computer vision [84, 218, 252].

Besides these four main types of explanations, there are other explanations such as decision rules [60, 206] or game strategies [92, 180] when AI plays games. Explanations for medical diagnosis systems are more complex. For instance in [181, 182], ontology-based explanations [182] are designed for the application called Doctor AI [47] (RNN-based model) specifically. As mentioned earlier, recommender systems represent a popular application domain to study explanations (marked with $*$), for instance [64, 82, 187, 188] use content-based explanations for recommendations. Besides, novel explanations such as dataset explanations [14] or explainable interfaces (marked with †) are used to study the effect of XAI. More details about concrete names of models and explanations can be found in Table 9 in Appendix.

		White-box	Black-box	Other
Feature-based	**local**	[2, 26, 40, 45, 115, 231, 243] [18]	[8, 93, 115, 169, 193, 219] [22, 26, 32, 50, 96, 244] [7, 18, 21, 42, 173, 218] [23, 96, 139, 140, 211]	
	global	[18, 63, 115, 185, 243]	[18]	
Example-based		[63, 115, 231, 243]	[31, 34, 37, 96, 140, 218]	[199] (generative models)
Counterfactual		[115, 130, 131, 243]	[209, 218, 244, 258]	
Concept-based			[21, 84, 142, 147, 252, 256] [218, 258]	
Other		[60, 92, 180, 206]	[34, 53, 181, 182, 225] [16, 42, 53, 173] [72, 143, 169, 191, 193] [64, 82, 187, 188, 241]* [128, 136, 175, 230]*	[14] [48, 148, 148, 159, 196, 229] † [17, 75, 106, 143, 149, 207, 220] †

Table 3. Models and explanations in core papers. 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.

4.2 Measurements

4.2.1 Trust. User trust is studied in decision-making applications such as image classification [34, 37], (review) deception detection [140] or loan approval [209]. Besides decision making, [64, 136, 149, 159, 175, 229] study user trust in the domain of recommendation systems ranging from movies [136], art paintings [64] to math exercises for school students [175]. Whether explainable machine learning models can increase user trust in the medical domain is studied in [181, 225, 231]. Moreover, Colley et al. [53] measure user trust in an autonomous driving application with and without explanations of the driving senses.

Trust measures used in much of the existing research can be divided into two groups: *self-reported* and *observed* trust [183]. 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 10 (Appendix), trust measures in these two groups are listed. The agreement rate of users with the model decisions is commonly used [140, 206, 225, 243] as a measure of observed trust. 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* [243]. Parallel to observed trust measurement, van der Waa et al. [235] 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. In self-reported measurements, researchers either utilize well-developed questionnaires or self-designed ones, with the exception of [70] which conducts a semi-structured interview to explore user opinions. Several works [34, 37, 64, 120, 159, 206, 209, 229, 231] propose their own questionnaires. Among these, a subgroup [34, 64, 120, 159, 229] 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 [175]. When deploying previously proposed questionnaires [14, 45, 53, 73, 92, 115, 136, 149, 175, 180], Trust in Automation [112] is the most commonly used one, in which the underlying constructs of trust between human and computerized systems are explored.

4.2.2 Fairness. Users' feedback on fairness of AI models has been studied predominantly in high-stakes applications such as granting loans [209] or granting bail for criminal offenders [63, 88, 94]. It is also considered in everyday use

Tasks	Tabular	Image/Video	Text	Other
forward simulation	[16, 26, 45, 50, 96, 169, 185, 193, 243]	[7, 31, 34, 42, 173, 193]	[18, 96]	[257] (Audio)
marginal feature effects	[2, 32, 45, 96, 243]		[96]	
manipulation / counterfactual sim.	[45]	[199, 243]	[18]	
feature importance	[26, 32, 243]	[218]		
failure prediction		[42, 243]		
relative simulation (selection)	[45, 50]			
other	[2] (mental model faithfulness)	[173] (class-wise acc.)		

Table 4. Works measuring objective understanding grouped by proxy task/data modality

cases such as news [188] and music [108] recommendations, or possible career suggestions [241], where a bias in the underlying system can be to the detriment of the user. Many authors examine decision-making systems that are already deployed in practice. For example, [63, 88, 94] investigate the fairness of COMPAS, a commercial criminal risk estimation tool that is currently used in the US to help make judicial bail decisions. Furthermore, Rader et al. [188] consider Facebook's News Feed algorithm. Htun et al. [108] contextualize their studies in the Spotify API for music playlist recommendations.

Since 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" [63], which can be answered on 5- or 7-point Likert scales [14, 63, 88, 94, 188, 209]. Among these works, an effective explanation is the one that can increase the human perceptions of fairness, except in [63]. The authors of this work find that models which give local explanations are rated the least fair because they reveal a potential model bias. The *decrease* of the fairness perceptions is thus viewed as effective, since the aim of explanations should be that the fairness is only rated highly if the underlying system is actually fair [63]. In addition to quantitative questions for fairness, in some studies participants are asked to give the reasoning behind their fairness ratings, either by choosing from a selection of statements [88] or by providing a written open text response [88, 209]. Moreover, Anik and Bunt [14] conduct a semi-structured interview after their questionnaire to capture reasoning and general perceptions of the given data-centric explanations.

4.2.3 Understanding. An important goal of explanation techniques is to foster users' understanding of complex machine learning systems. However, quantifying users' understanding remains challenging [154]. 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 [50, 96]. Cheng et al. [45] 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 [84, 119], where the intelligibility of concepts needs to be verified. We refer to this third category as *understanding of explanations* but defer detailed findings for these types of checks to Appendix C.3.

Objective Understanding. Works in the subdomain of objective understanding are concerned with measuring the extent to which a model is verifiably understood by a user. Therefore, proxy tasks are being solved by the users to assess their understanding of the model's inner workings. The most commonly considered domain in works on understanding is finance [2, 26, 32, 50, 96, 169, 185] followed by image classification [31, 34, 115]. Interesting applications include the work by Zhang and Lim [257] that uses audio data for vocal emotion recognition and Ross et al. [199] who consider generative models such as VAEs [122] and GANs [86].

One of the most critical design choices when assessing objective understanding is the selection of a suitable proxy task. Doshi-Velez and Kim [66] argue that the task should *"maintain the essence of the target application"* that is anticipated. One of the most prominent tasks is forward simulation [66, 154]. 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* [154]. However, scholars have designed many more tasks to quantify understanding and applied them across a variety of data modalities (cf. Table 4 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* [66] 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 [154] pointed out that simulatability can only be a reasonable measure, if the model is simple enough to be captured by humans. As this might not be the case in domains such as computer vision, simpler proxy tasks are required [218]. 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 are *marginal effects queries*, where the subjects predict how changes in a given input feature will affect the prediction (a typical query would be: Does increasing feature X lead to a higher prediction of y being class 1?). Because explanations should allow to identify weaknesses in models, the task of *failure prediction* measures the accuracy of users' prediction when the model prediction will be wrong.

Subjective Understanding. Fostering a solid understanding of a model's behavior and its results is a challenging task on a general as well as on a user-specific level. 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. We refer to the category of perceived measurements of understanding as subjective understanding in this work. The most commonly used applications that measure subjective understanding are various recommendation system setups [64, 93, 187, 188].

Most of the works assess the subjective understanding of a user with a post-task questionnaire. Guo et al. [92] adapted a popular questionnaire designed for recommendation systems by Knijnenburg et al. [124], while Bell et al. [26] accommodated the questionnaire which originally intended to measure the intelligibility of differenet explanations by Lim and Dey [152]. On the other hand, agreement to simple subjective statements such as *"I understand this decision algorithm"* [45], *"I understand how the AI..."* [34, 37] or *"The explanation(s) help me to understand..."* [187] is be collected to assess subjective understanding. Less directed questions can be e.g., *"Does this explanation show me why the system thought what it did?"* [96].

4.2.4 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 ranging from recommendation systems such as for art [64], music [128, 159], or a social context [229, 230] to decision making tasks such as house price prediction [2, 106]. We also include application-specific performance measures in this category. For instance, explanation studies in model debugging applications [21, 120, 185] are considered in this subsection.

Based on the measurements in the user studies, we refined usability into measures of helpfulness, workload (cognitive load), satisfaction, ease of use and performance in debugging applications, as shown in Table 1. To assess workload (cognitive load), NASA-TLX scale [95] is used in [17, 53, 64, 115, 231], while Abdul et al. [2] 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 [64, 128, 159, 219, 220, 229–231], however, the Explanation Satisfaction Scale [105] can be deployed as an established alternative [32, 181]. Helpfulness can be assessed by simply asking for subjective ratings of the explanations for accomplishing a specific task [34, 82, 173, 244, 256, 257]. Colley et al. [53] use an adapted version of the System Usability Scale by Holzinger et al. [107]. In debugging applications, the effectiveness of explanations is evaluated by objective performance measures, such as the number of bugs identified [21], the share of participants that identify a certain bias [218, First Experiment] or by the deviations between model predictions and human predictions for unusual samples [185]. An overview of used measures for usability can be found in Table 11 in Appendix.

4.2.5 Human-AI Team 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 [23]. Improving human performance with the help of AI has been considered in games [60, 180], question answering tasks [23, 75], deception detection [139, 140] and topic modeling [219, 220].

The most common assessment is to rate AI-aided human performance by the percentage of correctly predicted instances in the decision-making process (text classification) [23, 139, 140]. Paleja et al. [180], however, define the performance as the time to complete the task. In [60], 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. Additionally, the authors also measure perceived performance by a 5-point Likert scale rating of the question *"Do you believe your performance improved this session?"*.

4.3 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 5 and Figure 3, respectively.

	Experimental Design		
	Between-Subjects	Within-Subjects	Mixed
	[92, 149, 175, 209, 243]		
	[37, 45, 115, 136, 140]	[53, 70, 181, 225, 229]	
	[72, 96, 199, 206, 257]	[115, 120, 159, 180, 230]	[14, 17, 31, 34, 64]
Papers	[16, 18, 32, 169, 185]	[31, 34, 50, 75, 196]	[63, 72, 93, 142, 180]
	[26, 130, 173, 188, 191]	[21, 28, 81, 106, 218]	[187, 192, 207, 243, 244]
	[7, 8, 139, 219, 220]	[84, 128, 147, 231, 256]	[252]
	[42, 88, 188, 218, 243]	[82, 211]	
	[28, 48, 94, 108, 131, 241]		

Table 5. Experimental designs in core papers.

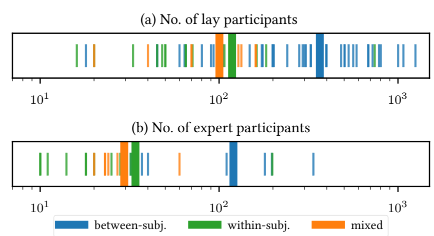

Fig. 3. Distribution of participant numbers in the surveyed user studies by design and participant type (each bar represents one study). Per-design means are indicated in bold.

Between-subjects. With slightly above 53 % 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. Experiments in a between-subjects manner usually start at around 30 participants, while the number of participants may go up to 1070 in total for 3 conditions used in [37] and to 1250 in 5 conditions in [185]. However, the number of participants can be limited such as in [175], which uses 12 per condition, as the studied model (application) is designed for school students and the participants cannot be easily recruited from the Internet platforms such as Amazon Mechanical Turk or Prolific. Some authors place particular emphasis on participants being similar to the average demographic [88, 94].

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 (e.g., the design by Abdul et al. [2] with 3 independent variables leads to 27 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 [14, 88, 209].

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 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 Tukey HSD test (e.g., [169]). When comparing individual samples, some works make use of the Bonferroni correction.

Within-subjects. Around 29 % 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. [225] and Rong et al. [196] 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 [196], however, it is often the case in expertise research. Gegenfurtner et al. [83] evaluate 73 sources and point out that the majority of these studies include only five, maybe ten experts. Besides the medical domain, other works [53, 70, 115, 229] also invite subjects with particular professions such as engineers in a technology company. When no specific knowledge is required, however, participant numbers reach up to 740 also for within-subjects designs [81].

For within-groups designs, the Wilcoxon signed-rank test (e.g. used by [31, 50]) 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., [50]).

Mixed. The smallest group of studies, about 18%, use a mixture of between- and within-subjects settings. In these works, subjects are first assigned randomly to one group, and inside each group, they are all exposed to multiple conditions. Anik and Bunt [14] use the knowledge backgrounds 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. [64] 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 the study design is that statistical tools from both the independent-variables and dependent-variables categories need to be combined.

5 FINDINGS OF USER STUDIES

In this section, we summarize the primary findings from the core papers. Table 6 and Table 7 list these findings with respect to five measured quantities. To build an overview 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 6. The listed impacts using explanations are to be seen in comparison with a control group without explanations. Beyond the explanations themselves, other possible evaluation dimensions that might might have an impact on the perception of XAI are in Table 7. For instance, these include AI technology literacy, model performance, or the dimensionality of the data. 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).

As various research questions and findings are addressed in 85 core papers, for instance, 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 Tables 6 and 7, 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 be validated by the number of papers obtaining positive or non-positive effects in each category, while the last findings can be found in Table 7. We encourage our readers 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 from the table for each category of measurement.

Trust. Among the papers comparing the effect of using explanations to using no explanations, or placebo (randomly generated) explanations [140, 175], about half of the papers validate that explanations have a positive impact on user trust [34, 64, 72, 140, 175, 180, 181, 209], while the other half cannot verify this hypothesis [45, 53, 115, 120, 206, 243]. For instance, Colley et al. [53] 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 [243], 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 [14, 37, 45, 136, 159, 220]. Millecamp et al. [159] 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 [149] 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 [37] 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).

Fairness. Overall, when compared to the control group with no explanations, few papers validate the effectiveness of explanations in improving users' perception of fairness. Only Schoeffer et al. [209] find out that showing feature importance scores or counterfactual explanations (or combination of both) for explaining decisions helps increase the feeling of fairness, where highlighting important features without scores does not. In [188], the authors use ranking

		Evaluation Dimension: Explanations Effect of explanations compared to **no explanations**	
		Positive	Non-positive / Mixed
Trust		[34]: example-based, rule-based explanations [64]: example-based explanations for recommendations [209]: feature importance [180]: decision-tree explanation for policy [72]: explanation corpus given by researchers [140]: feature-based (saliency map), example-based explanations [175]: explanations for recommendations [231]: rationale-based, example-based and feature-based (best) explanations for online symptom checkers	[53]: positive in simulation but no improvement in real-word [181]: explanations for medical suggestions (Doctor XAI [182]) pos.for observed trust but insignificant for reported trust [243]: feature-based explanations increase appropriate trust slightly but counterfactual explanations inconclusively [45, 115, 120]: feature-based explanation, insignificant [206]: rule-based explanation, insignificant [219]: feature-based explanation, negative
Fairness		[209]: feature importance, counterfactual explanations	[188]: informing users about the algorithmic decisions, negative ranking scores of recommendations, insignificant [209]: highlight features only, insignificant [28]: insignificant in between-subjects but significant in within-subjects
Understanding	Obj.	[45, 185] white-box model [96] feature importance, LIME (tabular) [257] counterfactuals+cues (audio) [18] manipulatability improved by white-box log. reg. [7] saliency maps (image) [243] counterfactuals+feature importance	[26]: SHAP, negative for black-box model (education domain) [26]: Insignificant difference btw. black-box and white-box models [96]: Prototypes, Anchors, LIME on textual data insignificant [257]: Counterfactuals and Concepts insignificant (audio data) [18]: Simulatability results insignificant for LIME, IG, surrogate model on BERT and Logistic Regression Model, Manipulatability insignificant for BERT [42]: Insignificant results for GRAD-CAM, Saliency Map, uncertainty scores in VQA
	Sub.	[34]: example-based, rule-based explanations [72]: explanation corpus given by researchers [243]: feature-, example- and counterfactual-based [188]: explanations provided by [55] for Facebook News Feed [64]: example- and feature-based explanations [37]: example-based explanations [93]: feature importance, SHAP and LIME [50]: feature importance, SHAP	[45]: white-box model, insignificant [26]: white-box < black-box, both insignificant
Usability		[106]: counterfactuals, pos. for usability [32, 64]: example-based explanations, pos. for satisfaction [256]: CAM-related explanations, pos. for helpfulness [231]: rational-, feature-, example-based explanations, pos. for satisfaction [230]: content-based explanations, pos. for satisfaction [207]: explanations regarding driving information, pos. for ease of use [34]: example-based and rule-based explanations, pos. for helpfulness [21]: local, global, visual (saliency map) explanations, pos. for bug identification [244]: attribution methods and conceptual explanations, pos. for usefulness [48]: feature-based, pos. for reliability [120]: (proposed) template-based expl. pos. for debugging and usefulness	[131]: counterfactuals, significant for helpfulness/usability but insignificant for usefulness [181]: ontology-based explanation, insignificant for satisfaction [244]: attribution methods and conceptual explanations, insignificant for ease of use [120]: visual explanations increases usefulness, but improvement is insignificant [53]: pos. for cognitive load/usability (simulation), but insignificant in real world [219]: feature-based explanations, negative for satisfaction
Human-AI Team Performance		[60]: textual explanations with domain knowledge (in chess) [139, 140]: feature-based explanations [75]: exampled-based for experts, feature-based for novices [81]: contrastive explanations [34]: example-based and rule-based explanations	[140]: exampled-based, insignificant [23]: feature-based explanations, insignificant

Table 6. User study findings when using model **explanations** as evaluation dimensions. 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.

scores of recommended items as explanations as well as giving information about other features in recommendation-making. These explanations do not have effects on user fairness perceptions. However, if users are informed about that the recommendations are made by algorithms, they reckon the system as unfair compared to only seeing the

		Other Evaluation Dimensions	
		Positive	Non-positive / Mixed
Trust		[14]: balanced training data, [37]: high model performance [136]: high quality of explanations [209]: high AI literacy [219]: interactivity	[14]: user expertise, insignificant [159]: personal characteristics, insignificant [220]: different topic modeling approaches, insignificant [149]: self-referential pronoun "I" in explanations, negative [45]: user technical literacy, insignificant
Fairness		[209]: high AI literacy [88]: fair features are “current charges” and “criminal history”	[88]: unfair features are “quality of school life” and “education & school behavior”, etc.
Understanding	Obj.	[199]: disentanglement of gen. model [45]: interactivity	[185, 199]: high dimensionality, negative [32]: contextualization, insignificant [34]: inductive vs. deductive explanations, insignificant [18]: different ML models, insignificant [31]: user expertise, insignificant [42]: instant feedback, insignificant [173]: timing of model errors, mixed
	Sub.	[199]: disentanglement of gen. model [45]: interactivity	[96]: model correctness, insignificant [191]: QuickSort, insignificant [50]: test of understanding, negative
Usability		[199]: significant difference in self-reported difficulty dependent on the generative model [92, 219]: interactivity [17]: Parallel Embeddings	[220]: different topic modeling approaches, insignificant [159]: personal characteristics, insignificant for satisfaction [173]: early encounters of system weaknesses lead to lower explanation usage [185]: clear model is *less* useful in debugging
Human-AI Team Performance		[139]: low model complexity [8, 34]: showing model prediction	[180]: explanations are positive for novices’ performance but negative for experts’

Table 7. User study findings when using model **others** (other than explanations) as evaluation dimensions. Effects on measured quantities are divided into "Positive" where explanation information is given, and "Non-positive / Mixed" where negative impact is marked with underlines.

suggestions. Contradictory results are found in [28] regarding between-subjects and within-subjects designs: Users in within-subject designs have significantly different impressions of explanations for one single case, where case-based explanations have negative impact. Surprisingly, in between-subjects designs, where users are exposed to multiple cases explained by only one explanation technique, the differences among explanation methods disappear. Both scenarios (viewing multiple explanations for one decision or viewing various decisions explained by one explanation style) can happen in the real-world [28], therefore the effectiveness of explanations in fairness perception is highly dependent. Besides explanation styles, AI literacy of participants is studied in [209], where a positive correlation between fairness and AI literacy is observed. Grgić-Hlača et al. [88] let users assign fairness scores for features and some features such as “current charges” are rated as fair and some such as “education & school behavior” are unfair.

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. [7] who reported that saliency maps have a positive effect on understanding. Wang and Yin [243] show that counterfactual explanations and feature importance increase users objective understanding. On the contrary, Sixt et al. [218] find none of their examined explanation techniques (counterfactuals, conceptual explanations) superior to a baseline technique consisting of example images for each class and Hase and Bansal [96] reveal 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 [96] 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. [18] show that their manipulatablity task revealed differences that remained hidden when forward simulation was used. In spite of these findings, Buçinca et al. [34] underline that the preferred explanation may be different in a real decision making task than in the simulation. 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 explanation resulted in higher simulation performance than white-box models with SHAP values in [26], Cheng et al. [45] 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 [37, 45, 92, 93, 130, 173, 187, 199, 243]. There exist a trend of using model explanations to improve subjective understanding [34, 37, 64, 72, 93, 188, 242]. However, Chromik et al. [50] challenge the improvement in perceived understanding with the cognitive bias named *illusion of explanatory depth* (IOED) [200], 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 [45], where white-box explanations increase objective understanding but do not have have significant impact on subjective understanding. Similar disagreements have been observed in multiple other works [96, 242]. Radensky et al. [187] examine the joint effects of local and global explanations in a recommendation system and their results provide evidence that both are better for the users' perceived understanding than either alone.

Usability. Similar to trust, it is not clear whether explanations are effective in improving users' perceptions in helpfulness, satisfaction or other dimensions in usability. For instance, in [32, 64, 220], the explanations have a positive effect on satisfaction, while no any significant effects on satisfaction are observed in [128, 159, 219, 229]. Parallel to trust, Smith-Renner et al. [219] provide evidence for the hypothesis that it is harmful for user trust and satisfaction to show explanations by highlighting the important words in a text classification task. A strong correlation in self-reported trust and satisfaction can also be observed in [53], where explanations have positive impact in simulated driving environment, but no significant effects when using real-world data. Beyond explanations, Nourani et al. [173] study the order of observing system weakness and strengths, which reveals that encountering weakness first results in lower rate of usage of system explanations than encountering strength first. Effects of explanations may be dependent on input samples, as shown in [256]. 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.

Human-AI Team Performance. A strain of works [60, 75, 139, 140] show that viewing explanations can improve human accuracy in making decisions, especially with feature-based explanations taking text data as input [75, 139, 140]. When using example-based explanations in text classification, there is no improvement in human performance [140]. Likewise, utilizing explanations has no significant impact on human performance in [8, 23], but simply showing model predictions has a positive effect in [8]. Experts and novices perceive explanations differently, for example, Feng and Boyd-Graber

[75] conclude that the performance gain of novices and experts comes from different explanation sources. Paleja et al. [180] reveal that explanations can improve novices' performance but decrease experts' performance. Additionally, less complex models with explanations can better convince humans in correct decisions [139].

6 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 guideline 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" [14, 53, 115] or "general trust in technology" [92, 136]. To further construct "trust" as a quantitative measurement, one needs to examine how existing work have conceptualized "trust" in both social sciences context as well as XAI and technical context [104]. Application-based quantity depends on the application goal, for instance in a chess game [60], the measurement is the human winning percentage with the help of model explanations (Human-AI team performance).

From Table 6, 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 Table 12 in Appendix). 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 [72, 84, 211]).

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 [7]. Thus other proxy tasks such as feature importance queries [218] or manipulatability checks [18, 199] were proposed. Another important point is to choose a proxy task that is simplified, but features many characteristics of the application in mind [66]. 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 [34].

The measurement is often dependent on the definition of the measured quantity. For instance, in [42], the objective understanding is measured as failure prediction (the accuracy of users 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 it cannot completely reflect different constructs of measured quantities [175]. 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 [206]. Similar findings have been made with respect to objective and subjective understanding [50, 96, 243]. To overcome this limitation, both self-reported and observed measures shall be used in parallel.

Besides the measures introduced in Section 4.2, there are several psychological constructs which 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 [69], which is not thoroughly studied in the XAI experience yet. The subjective task value consists of intrinsic value (enjoyment),

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

attainment value (importance for one's self), utility value (usefulness), and cost (the amount of effort or time needed) [69, 111]. 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 [95, 179]. 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 [176]. 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[2] has become a common practice in recent years [215]. 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 [216] and thus strengthen the credibility of a study. Expert interviews and pre-studies following a think-aloud protocol [74], e.g., in the references [199, 257], 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 [34, 244].

During the User Study. 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 reminder the day before and provide them with your contact in case they could not 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 should participants leave their backpack, water bottle, and lunch box) and plans for unexpected situations (e.g., uncooperative participants and malfunctional 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 [104]. 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.

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 Figure 3 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 [52]. 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 models explanations need to evolve as well [62].

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. [135] justify the

[2]https://aspredicted.org

use of these checks without compromising scale validity. In within-subject experiments, a random order of conditions is necessary to avoid order effect [181]. Participants can learn knowledge of data or examples shown in the previous conditions, and Tsai et al. [231] choose to use a Latin square design to avoid the learning effect.

After the User Study. After the data collection is finished, 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 4.3. 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 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 [78] 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 to assess the the characteristics 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).

7 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.

7.1 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 development of XAI solutions that better respond to user needs. As discussed in [195], there are two aspects in human-centered AI: (1) AI systems that understand humans with the sociocultural background and (2) AI systems that help humans understand them. The former point can be used to guide the design of AI systems. In this section, we discuss XAI research that leverages this insight.

The process of explaining 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 on how humans perceive other intelligent agents (humans, animals, 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) 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 explainable and transparent AI systems and human-centered evaluations.

Expectancy-value Motivation Theory. The question of how to enhance the efficiency and the outcome of this human learning process is of high importance [137]. This research question is widely considered in educational psychology through the lens of expectancy-value motivation theory [111, 194, 248]. For instance, Hulleman et al. [111] 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 during 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 [46, 98]. Bisra et al. [29] summarize guidelines for effective scaffolding. For instance, different disciplinary descriptions can be used in the scaffolding (explanation prompt) to enhance the user's intuition. Another important, yet often unconsidered point is the role of personality traits in the perception of explanations. For instance, Conati et al. [54] show that the *need for cognition* characteristic, which indicates users' openness towards cognitively challenging tasks, is a determining factor for explanation effectiveness in an intelligent tutoring system. Considering these findings, we see personalized XAI as a relatively underexplored but yet sorely needed research direction.

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 Theory of Mind (ToM) [24, 25] in psychology. Based on this theory, Bayesian Theory of Mind (BToM) [20] provides a probabilistic framework to predict the inferences that people make about the mental states underlying other agents' actions [57]. Recent work, at the intersection of XAI and robotics, indicates that humans also attribute ToM to artificial agents that they observe or interact with [99, 141]. Guided by these user-centered results, several works also at the intersection of XAI and robotics have utilized BToM to create a simulated user, then use the simulated user to generate helpful explanations. Towards this goal, Huang et al. [110] provide a greedy algorithm for selecting explanations that maximize the simulated user's knowledge of the agent's (a self-driving car in their domain) policy; and Lee et al. [146] provide a related approach where the user is modeled as an inverse reinforcement learner. In addition to selecting the most informative explanations, Qian and Unhelkar [186] utilize a variation of the Monte Carlo tree search to generate a computationally tractable approach to identify the most informative sequence of the explanations, based on the assumption that some explanations might be more effective initially. Thus, while some existing works evaluate the effectiveness of the selected explanations through experiments with human users, the community still lacks understanding of how robust or realistic BToM is compared to a human's cognitive process particularly for XAI. We also advocate for more probabilistic and computational cognitive models to be utilized in XAI designs. To achieve this, we need experts from cross disciplines to address individual user's needs in an XAI system from cognitive, psychological, and computational perspectives. Lastly, we also encourage XAI researchers to develop solutions to explain *AI-enabled systems* – for instance, robots and autonomous vehicles – which require grounded and user-centered solutions.

Hybrid Teaching. Teaching strategies for the human-to-human setting have been widely studied and many categorizations exist [114, 172, 202]. 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 [11, 12, 109, 146, 245, 255]. 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. Although the goal of direct and indirect teaching methods is the same, namely explaining an AI system to human users, the computational problems solved by these methods are different. 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 [80, 165, 166]. In XAI related work, Qian and Unhelkar [186] provide a hybrid teaching framework by introducing an *AI Teacher* to enable guided interactivity between RL-based AI agents and a user. Their results indicate that hybrid teaching reduce 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 or model enabling active learning.

7.2 Open Research Problems

Automatic evaluations vs. 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 [66, 170]. These metrics aim to test desiderata around the "faithfulness"/"fidelity"/"truthfulness" of model explanations [9, 170, 228]. Faithfulness of explanations is defined as that explanations are indicative of true important features in the input [9]. 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 examine that 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 in essential to guarantee the validity of the empirical user study and make 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.

XAI should be calibrated to account for model quality. As suggested in the last point, a good model explanation should be faithful to the model, therefore, it should reveal the weaknesses of the model. When seeing unexpected explanations, users will express their negative feelings about the model explanations and the explanations are rated as bad. Nevertheless, these "bad" explanations might truly reflect the imperfections of models and the low rating could be a sign of the usefulness of model explanations. Researchers in the XAI domain consider this dilemma in user trust evaluation but sporadically in other quantity assessments. Good model explanations should help users *calibrate* their trust [36, 196], i.e., trust the model decision when it is correct but distrust it when incorrect. In fact, there is no conciseness that users' giving positive feedback upon an evaluation quantity indicates the effectiveness of explanations. For instance, when evaluating model fairness, [14, 88, 94, 188, 209] reckon the increase in perceived fairness as positive, while Dodge et al. [63] define the decrease as positive. Therefore, Schoeffer and Kuehl [208] propose a novel desideratum, *appropriate fairness perceptions*, meaning that people increase or decrease their fairness perceptions if the model is fair or biased. For future work, more attention should be paid to the evaluation of *calibrated* perceptions and make sure that model explanations can faithfully convey model information.

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 [66];
(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. [34] 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 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.

Simulated evaluation as a cost-efficient solution? As human-subject experiments are costly to conduct, Chen et al. [44] 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 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 [5]. Undoubtedly and also affirmed by Chen et al. [44], 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 7.1 discusses several candidate human cognition models and highlights recent XAI works [137, 186] that utilize this "Oz-of-Wizard" paradigm [223].

8 CONCLUSION

In recent years, there has been a proliferation of XAI research in both academia and industry. Explainability is a human-centric property [151] 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] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–18.

[2] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y Lim. 2020. COGAM: measuring and moderating cognitive load in machine learning model explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[3] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[4] Peter Martey Addo, Dominique Guegan, and Bertrand Hassani. 2018. Credit risk analysis using machine and deep learning models. *Risks* 6, 2 (2018), 38.

[5] Gati Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2022. Using Large Language Models to Simulate Multiple Humans. *arXiv preprint arXiv:2208.10264* (2022).

[6] Maher Ala'raj, Maysam F Abbod, Munir Majdalawieh, and Luay Jum'a. 2022. A deep learning model for behavioural credit scoring in banks. *Neural Computing and Applications* 34, 8 (2022), 5839–5866.

[7] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 275–285.

[8] Yasmeen Alufaisan, Laura R Marusich, Jonathan Z Bakdash, Yan Zhou, and Murat Kantarcioglu. 2021. Does explainable artificial intelligence improve human decision-making?. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 6618–6626.

[9] David Alvarez Melis and Tommi Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems* 31 (2018).

[10] Filippo Amato, Alberto López, Eladia María Peña-Méndez, Petr Vaňhara, Aleš Hampl, and Josef Havel. 2013. Artificial neural networks in medical diagnosis. , 47–58 pages.

[11] Dan Amir and Ofra Amir. 2018. HIGHLIGHTS: Summarizing Agent Behavior to People. In *AAMAS*.

[12] Ofra Amir, Finale Doshi-Velez, and David Sarne. 2019. Summarizing agent strategies. *Autonomous Agents and Multi-Agent Systems* 33, 5 (2019), 628–644.

[13] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. In *Ethics of Data and Analytics*. Auerbach Publications, 254–264.

[14] Ariful Islam Anik and Andrea Bunt. 2021. Data-centric explanations: explaining training data of machine learning systems to promote transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[15] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*. International Foundation for Autonomous Agents and Multiagent Systems, 1078–1088.

[16] Javier Antoran, Umang Bhatt, Tameem Adel, Adrian Weller, and José Miguel Hernández-Lobato. 2021. Getting a {CLUE}: A Method for Explaining Uncertainty Estimates. In *International Conference on Learning Representations*.

[17] Dustin L Arendt, Nasheen Nur, Zhuanyi Huang, Gabriel Fair, and Wenwen Dou. 2020. Parallel embeddings: a visualization technique for contrasting learned representations. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 259–274.

[18] Siddhant Arora, Danish Pruthi, Norman Sadeh, William W Cohen, Zachary C Lipton, and Graham Neubig. 2022. Explain, edit, and understand: Rethinking user study design for evaluating model explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 5277–5285.

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

[20] Chris Lawrence Baker. 2012. *Bayesian Theory of Mind: Modeling Human Reasoning about Beliefs, Desires, Goals, and Social Relations*. Ph. D. Dissertation. Massachusetts Institute of Technology, Dept. of Brain and Cognitive Sciences.

[21] Agathe Balayn, Natasa Rikalo, Christoph Lofi, Jie Yang, and Alessandro Bozzon. 2022. How can Explainability Methods be Used to Support Bug Identification in Computer Vision Models?. In *CHI Conference on Human Factors in Computing Systems*. 1–16.

[22] Seojin Bang, Pengtao Xie, Heewook Lee, Wei Wu, and Eric Xing. 2021. Explaining a black-box by using a deep variational information bottleneck approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11396–11404.

[23] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[24] S. Baron-Cohen. 1991. Precursors to a theory of mind: Understanding attention in others. *Whiten, Andrew (ed.), Natural theories of mind* (1991), 233–251.

[25] G. S. Becker. 1976. *The economic approach to human behavior*. University of Chicago press.

[26] Andrew Bell, Ian Solano-Kamaiko, Oded Nov, and Julia Stoyanovich. 2022. It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 248–266.

[27] Izak Benbasat and Weiquan Wang. 2005. Trust in and adoption of online recommendation agents. *Journal of the association for information systems* 6, 3 (2005), 4.

[28] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*. 1–14.

[29] Kiran Bisra, Qing Liu, John C Nesbit, Farimah Salimi, and Philip H Winne. 2018. Inducing self-explanation: A meta-analysis. *Educational Psychology Review* 30, 3 (2018), 703–725.

[30] Miranda Bogen and Aaron Rieke. 2018. Help wanted: An examination of hiring algorithms, equity, and bias. *Upturn, December* 7 (2018).

[31] Judy Borowski, Roland Simon Zimmermann, Judith Schepers, Robert Geirhos, Thomas S. A. Wallis, Matthias Bethge, and Wieland Brendel. 2021. Exemplary Natural Images Explain {CNN} Activations Better than State-of-the-Art Feature Visualization. In *International Conference on Learning Representations*.

[32] Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2022. Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users. In *27th International Conference on Intelligent User Interfaces*. 807–819.

[33] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[34] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th international conference on intelligent user interfaces*. 454–464.

[35] Nadia Burkart and Marco F Huber. 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research* 70 (2021), 245–317.

[36] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*. IEEE, 160–169.

[37] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces*. 258–262.

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

[39] Francisco Maria Calisto, Carlos Santiago, Nuno Nunes, and Jacinto C Nascimento. 2022. BreastScreening-AI: Evaluating medical intelligent agents for human-AI interactions. *Artificial Intelligence in Medicine* 127 (2022), 102285.

[40] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 1721–1730.

[41] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832.

[42] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. 2018. Do explanations make VQA models more predictable to a human?. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 1036—-1042.

[43] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. 2015. Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE international conference on computer vision*. 2722–2730.

[44] Valerie Chen, Nari Johnson, Nicholay Topin, Gregory Plumb, and Ameet Talwalkar. 2022. Use-Case-Grounded Simulations for Explanation Evaluation. *arXiv preprint arXiv:2206.02256* (2022).

[45] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[46] Olga Chernikova, Nicole Heitzmann, Ansgar Opitz, Tina Seidel, and Frank Fischer. 2022. A Theoretical Framework for Fostering Diagnostic Competences with Simulations in Higher Education. *Learning to Diagnose with Simulations* (2022), 5.

[47] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*. PMLR, 301–318.

[48] Saemi Choi, Kiyoharu Aizawa, and Nicu Sebe. 2018. Fontmatcher: font image paring for harmonious digital graphic design. In *23rd International Conference on Intelligent User Interfaces*. 37–41.

[49] Steven A. Cholewiak, Panos Ipeirotis, Victor Silva, and Arun Kannawadi. 2021. *SCHOLARLY: Simple access to Google Scholar authors and citation using Python*. https://doi.org/10.5281/zenodo.5764801

[50] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I think i get your point, AI! the illusion of explanatory depth in explainable AI. In *26th International Conference on Intelligent User Interfaces*. 307–317.

[51] Michael Chromik and Martin Schuessler. 2020. A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI. *Exss-atec@ iui* 94 (2020).

[52] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Routledge.

[53] Mark Colley, Benjamin Eder, Jan Ole Rixen, and Enrico Rukzio. 2021. Effects of semantic segmentation visualization on trust, situation awareness, and cognitive load in highly automated vehicles. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–11.

[54] Cristina Conati, Oswald Barral, Vanessa Putnam, and Lea Rieger. 2021. Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence* 298 (2021), 103503.

[55] Kelley Cotter, Janghee Cho, and Emilee Rader. 2017. Explaining the news feed algorithm: An analysis of the" News Feed FYI" blog. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*. 1553–1560.

[56] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-adapted interaction* 18, 5 (2008), 455–496.

[57] Gergely Csibra. 2017. Cognitive science: Modelling theory of mind. *Nature Human Behaviour* 1 (2017). Issue 4. https://doi.org/10.1038/s41562-017-0066

[58] Henggang Cui, Vladan Radosavljevic, Fang-Chieh Chou, Tsung-Han Lin, Thi Nguyen, Tzu-Kuo Huang, Jeff Schneider, and Nemanja Djuric. 2019. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2090–2096.

[59] Arun Das and Paul Rad. 2020. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371* (2020).

[60] Devleena Das and Sonia Chernova. 2020. Leveraging rationales to improve human task performance. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 510–518.

[61] Xolani Dastile, Turgay Celik, and Moshe Potsane. 2020. Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing* 91 (2020), 106263.

[62] Shipi Dhanorkar, Christine T Wolf, Kun Qian, Anbang Xu, Lucian Popa, and Yunyao Li. 2021. Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In *Designing Interactive Systems Conference 2021*. 1591–1602.

[63] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*. 275–285.

[64] Vicente Dominguez, Pablo Messina, Ivania Donoso-Guzmán, and Denis Parra. 2019. The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 408–416.

[65] Tim Donkers, Timm Kleemann, and Jürgen Ziegler. 2020. Explaining recommendations by means of aspect-based transparent memories. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 166–176.

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

[67] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaao5580.

[68] Natasha Dwyer, Hector Miller-Bakewell, Tessa Darbyshire, Anirban Basu, and Steve Marsh. 2021. The ethical implications of using digital traces: studying explainability and trust during a pandemic. In *Qualitative and Digital Research in Times of Crisis: Methods, Reflexivity, and Ethics*. Policy

Press, 129–142.
[69] Jacquelynne Eccles. 1983. Expectancies, values and academic behaviors. *Achievement and achievement motives* (1983).
[70] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.
[71] Upol Ehsan and Mark O Riedl. 2020. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *International Conference on Human-Computer Interaction*. Springer, 449–466.
[72] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 263–274.
[73] Bradley J Erickson, Panagiotis Korfiatis, Zeynettin Akkus, and Timothy L Kline. 2017. Machine learning for medical imaging. *Radiographics* 37, 2 (2017), 505.
[74] K Anders Ericsson and Herbert A Simon. 1984. *Protocol analysis: Verbal reports as data.* MIT Press.
[75] Shi Feng and Jordan Boyd-Graber. 2019. What can ai do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 229–239.
[76] Davide Ferrari, Giovanni Guaraldi, Federica Mandreoli, Riccardo Martoglia, Jovana Milić, and Paolo Missier. 2020. Data-driven vs knowledge-driven inference of health outcomes in the ageing population: a case study. In *Workshops of the 23rd International Conference on Extending Database Technology/23rd International Conference on Database Theory, EDBT-ICDT-WS 2020*, Vol. 2578. CEUR-WS.
[77] Juliana J Ferreira and Mateus S Monteiro. 2020. What are people doing about XAI user experience? A survey on AI explainability research and practice. In *International Conference on Human-Computer Interaction*. Springer, 56–73.
[78] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.
[79] Gerhard Friedrich and Markus Zanker. 2011. A taxonomy for generating explanations in recommender systems. *AI Magazine* 32, 3 (2011), 90–98.
[80] Catherine P. Fulford and Shuqiang Zhang. 1993. Perceptions of Interaction: The critical predictor in distance education. *American Journal of Distance Education* 7, 3 (1993), 8–21. https://doi.org/10.1080/08923649309526830
[81] Krzysztof Z Gajos and Lena Mamykina. 2022. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In *27th International Conference on Intelligent User Interfaces*. 794–806.
[82] Jingyue Gao, Xiting Wang, Yasha Wang, and Xing Xie. 2019. Explainable recommendation through attentive multi-view learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3622–3629.
[83] Andreas Gegenfurtner, Erno Lehtinen, and Roger Säljö. 2011. Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educational psychology review* 23, 4 (2011), 523–552.
[84] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. 2019. Towards Automatic Concept-based Explanations. In *Advances in Neural Information Processing Systems*, Vol. 32. 9277–9286.
[85] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.
[86] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).
[87] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine* 38, 3 (2017), 50–57.
[88] Nina Grgić-Hlača, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference*. 903–912.
[89] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. 2020. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics* 37, 3 (2020), 362–386.
[90] Maarten Grootendorst. 2020. KeyBERT: Minimal keyword extraction with BERT. https://doi.org/10.5281/zenodo.4461265
[91] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.
[92] Lijie Guo, Elizabeth M Daly, Oznur Alkan, Massimiliano Mattetti, Owen Cornec, and Bart Knijnenburg. 2022. Building Trust in Interactive Machine Learning via User Contributed Interpretable Rules. In *27th International Conference on Intelligent User Interfaces*. 537–548.
[93] Sophia Hadash, Martijn C Willemsen, Chris Snijders, and Wijnand A IJsselsteijn. 2022. Improving understandability of feature contributions in model-agnostic explainable AI tools. In *CHI Conference on Human Factors in Computing Systems*. 1–9.
[94] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. 2020. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 392–402.
[95] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.
[96] Peter Hase and Mohit Bansal. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5540–5552.
[97] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

[98] Nicole Heitzman, Tina Seidel, Ansgar Opitz, Andreas Hetmanek, Christof Wecker, Martin Fischer, Stefan Ufer, Ralf Schmidmaier, Birgit Neuhaus, Matthias Siebeck, et al. 2019. Facilitating diagnostic competences in simulations: A conceptual framework and a research agenda for medical and teacher education. *Frontline Learning Research* 7, 4 (2019), 1–24.

[99] Thomas Hellström and Suna Bensch. 2018. Understandable Robots. *Paladyn, Journal of Behavioral Robotics* 9 (07 2018), 110–123. https://doi.org/10.1515/pjbr-2018-0009

[100] Monika Hengstler, Ellen Enkel, and Selina Duelli. 2016. Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change* 105 (2016), 105–120.

[101] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. 241–250.

[102] Charlie Hewitt, Ioannis Politis, Theocharis Amanatidis, and Advait Sarkar. 2019. Assessing Public Perception of Self-Driving Cars: the Autonomous Vehicle Acceptance Model. In *Conference on Intelligent User Interfaces (IUI'19)*.

[103] Guy Hoffman. 2019. Evaluating fluency in human–robot collaboration. *IEEE Transactions on Human-Machine Systems* 49, 3 (2019), 209–218.

[104] Guy Hoffman and Xuan Zhao. 2020. A primer for conducting experiments in human–robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 1 (2020), 1–31.

[105] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[106] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[107] Andreas Holzinger, André Carrington, and Heimo Müller. 2020. Measuring the quality of explanations: the system causability scale (SCS). *KI-Künstliche Intelligenz* 34, 2 (2020), 193–198.

[108] Nyi Nyi Htun, Elisa Lecluse, and Katrien Verbert. 2021. Perception of fairness in group music recommender systems. In *26th International Conference on Intelligent User Interfaces*. 302–306.

[109] Sandy H. Huang, Kush Bhatia, P. Abbeel, and Anca D. Dragan. 2018. Establishing Appropriate Trust via Critical States. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2018), 3929–3936.

[110] Sandy H. Huang, David Held, Pieter Abbeel, and Anca D. Dragan. 2019. Enabling Robots to Communicate their Objectives. *Autonomous Robots* 43, 2 (February 2019).

[111] Chris S Hulleman, Jeff J Kosovich, Kenneth E Barron, and David B Daniel. 2017. Making connections: Replicating and extending the utility value intervention in the classroom. *Journal of Educational Psychology* 109, 3 (2017), 387.

[112] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics* 4, 1 (2000), 53–71.

[113] Gargi Joshi, Rahee Walambe, and Ketan Kotecha. 2021. A review on explainability in multimodal deep neural nets. *IEEE Access* 9 (2021), 59800–59821.

[114] Lynn Julien-Schultz, Nancy Maynes, and Cilla Dunn. 2010. Managing Direct and Indirect Instruction: A Visual Model to Support Lesson Planning in Pre-Service Programs. *The International Journal of Learning: Annual Review* 17 (2010), 125–140.

[115] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[116] Eoin M Kenny, Courtney Ford, Molly Quinn, and Mark T Keane. 2021. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence* 294 (2021), 103459.

[117] Anjali Khurana, Parsa Alamzadeh, and Parmit K Chilana. 2021. ChatrEx: Designing explainable chatbot interfaces for enhancing usefulness, transparency, and trust. In *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 1–11.

[118] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems* 29 (2016).

[119] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*. PMLR, 2668–2677.

[120] Dae Hyun Kim, Enamul Hoque, and Maneesh Agrawala. 2020. Answering questions about charts and generating visual explanations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[121] Jae Kyeong Kim, Hyea Kyeong Kim, Hee Young Oh, and Young U Ryu. 2010. A group recommendation system for online communities. *International journal of information management* 30, 3 (2010), 212–219.

[122] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[123] René F Kizilcec. 2016. How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2390–2395.

[124] Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User modeling and user-adapted interaction* 22, 4 (2012), 441–504.

[125] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*. PMLR, 1885–1894.

[126] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *International Conference on Machine Learning*. PMLR, 5338–5348.

[127] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2017. User preferences for hybrid explanations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 84–88.

[128] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 379–390.

[129] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5686–5697.

[130] Ulrike Kuhl, André Artelt, and Barbara Hammer. 2022. Keep Your Friends Close and Your Counterfactuals Closer: Improved Learning From Closest Rather Than Plausible Counterfactual Explanations in an Abstract Setting. *arXiv preprint arXiv:2205.05515* (2022).

[131] Ulrike Kuhl, André Artelt, and Barbara Hammer. 2022. Let's Go to the Alien Zoo: Introducing an Experimental Framework to Study Usability of Counterfactual Explanations for Machine Learning. *arXiv preprint arXiv:2205.03398* (2022).

[132] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.

[133] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1–10.

[134] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*. IEEE, 3–10.

[135] Franki YH Kung, Navio Kwok, and Douglas J Brown. 2018. Are attention check questions a threat to scale validity? *Applied Psychology* 67, 2 (2018), 264–283.

[136] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[137] Isaac Lage, Daphna Lifschitz, Finale Doshi-Velez, and Ofra Amir. 2019. Exploring computational user models for agent policy summarization. In *IJCAI: proceedings of the conference*, Vol. 28. NIH Public Access, 1401.

[138] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471* (2021).

[139] Vivian Lai, Han Liu, and Chenhao Tan. 2020. " Why is' Chicago'deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[140] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*. 29–38.

[141] Sau lai Lee, Ivy Yee man Lau, S. Kiesler, and Chi-Yue Chiu. 2005. Human Mental Models of Humanoid Robots. In *2005 IEEE International Conference on Robotics and Automation*. 2762–2772.

[142] Iro Laina, Ruth Fong, and Andrea Vedaldi. 2020. Quantifying Learnability and Describability of Visual Concepts Emerging in Representation Learning. *Advances in Neural Information Processing Systems* 33.

[143] Pierre Le Bras, David A Robb, Thomas S Methven, Stefano Padilla, and Mike J Chantler. 2018. Improving user confidence in concept maps: Exploring data driven explanations. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[144] John Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10 (1992), 1243–1270.

[145] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[146] Michael S. Lee, Henny Admoni, and Reid Simmons. 2021. Machine Teaching for Human Inverse Reinforcement Learning. *Frontiers in Robotics and AI* 8 (2021), 188. https://doi.org/10.3389/frobt.2021.693050

[147] Tobias Leemann, Yao Rong, Stefan Kraft, Enkelejda Kasneci, and Gjergji Kasneci. 2022. Coherence Evaluation of Visual Concepts With Objects and Language. In *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality*.

[148] Tianyi Li, Gregorio Convertino, Ranjeet Kumar Tayi, and Shima Kazerooni. 2019. What data should i protect? Recommender and planning support for data security analysts. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 286–297.

[149] Mengqi Liao and S Shyam Sundar. 2021. How Should AI Systems Talk to Users when Collecting their Personal Information? Effects of Role Framing and Self-Referencing on Human-AI Interaction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[150] Q Vera Liao, Milena Pribić, Jaesik Han, Sarah Miller, and Daby Sow. 2021. Question-driven design process for explainable ai user experiences. *arXiv preprint arXiv:2104.03483* (2021).

[151] Q Vera Liao and Kush R Varshney. 2021. Human-centered explainable AI (XAI): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790* (2021).

[152] Brian Y Lim and Anind K Dey. 2009. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th international conference on Ubiquitous computing*. 195–204.

[153] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2119–2128.

[154] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[155] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

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

[157] Saleem A Masadeh. 2021. *Investigating the YouTube Adpocalypse and Augmenting the User Experience to Address Transparency and Systemic Biases*. Ph. D. Dissertation. New Mexico State University.

[158] D Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and validating trust measures for e-commerce: An integrative typology. *Information systems research* 13, 3 (2002), 334–359.

[159] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. 2019. To explain or not to explain: the effects of personal characteristics when explaining music recommendations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 397–407.

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

[161] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11, 3-4 (2021), 1–45.

[162] Christoph Molnar. 2019. *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/.

[163] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2018. Methods for interpreting and understanding deep neural networks. *Digital signal processing* 73 (2018), 1–15.

[164] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. 2020. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter* 22, 1 (2020), 18–33.

[165] Brent Muirhead. 2001. Interactivity Research Studies. *Journal of Educational Technology & Society* 4, 3 (2001), 108–112. http://www.jstor.org/stable/jeductechsoci.4.3.108

[166] Brent Muirhead and Charles Juwah. 2004. Interactivity in computer-mediated college and university education: A recent review of the literature. *Educational Technology & Society* 7 (01 2004), 12–20.

[167] Robin R Murphy. 2019. *Introduction to AI robotics*. MIT press.

[168] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682* (2018).

[169] Karthikeyan Natesan Ramamurthy, Bhanukiran Vinzamuri, Yunfeng Zhang, and Amit Dhurandhar. 2020. Model agnostic multilevel explanations. *Advances in neural information processing systems* 33 (2020), 5968–5979.

[170] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2022. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *arXiv preprint arXiv:2201.08164* (2022).

[171] Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 1069–1078.

[172] Kevin A. Nguyen, Jenefer Husman, M. A. Trujillo Borrego, Prateek Shekhar, Michael J. Prince, Matt DeMonbrun, Cynthia J. Finelli, Charles Henderson, and Cindy K. Waters. 2017. Students' Expectations, Types of Instruction, and Instructor Strategies Predicting Student Response to Active Learning. *International Journal of Engineering Education* 33 (2017), 2–18.

[173] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *26th International Conference on Intelligent User Interfaces*. 340–350.

[174] Ingrid Nunes and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction* 27, 3 (2017), 393–444.

[175] Jeroen Ooge, Shotallo Kato, and Katrien Verbert. 2022. Explaining Recommendations in E-Learning: Effects on Adolescents' Trust. In *27th International Conference on Intelligent User Interfaces*. 93–105.

[176] Kim Ouwehand, Avalon van der Kroef, Jacqueline Wong, and Fred Paas. 2021. Measuring cognitive load: Are there more valid alternatives to likert rating scales?. In *Frontiers in Education*, Vol. 6. Frontiers Media SA, 702616.

[177] Ahmet Murat Ozbayoglu, Mehmet Ugur Gudelek, and Omer Berat Sezer. 2020. Deep learning for financial applications: A survey. *Applied Soft Computing* 93 (2020), 106384.

[178] Heather L O'Brien, Paul Cairns, and Mark Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies* 112 (2018), 28–39.

[179] Fred GWC Paas. 1992. Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. *Journal of educational psychology* 84, 4 (1992), 429.

[180] Rohan Paleja, Muyleng Ghuy, Nadun Ranawaka Arachchige, Reed Jensen, and Matthew Gombolay. 2021. The utility of explainable ai in ad hoc human-machine teaming. *Advances in Neural Information Processing Systems* 34 (2021), 610–623.

[181] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. 2022. Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems. In *CHI Conference on Human Factors in Computing Systems*. 1–9.

[182] Cecilia Panigutti, Alan Perotti, and Dino Pedreschi. 2020. Doctor XAI: an ontology-based approach to black-box sequential data classification explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 629–639.

[183] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. 2019. How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652* (2019).

[184] Seong Ho Park and Kyunghwa Han. 2018. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* 286, 3 (2018), 800–809.

[185] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.

[186] Peizhu Qian and Vaibhav Unhelkar. 2022. Evaluating the Role of Interactivity on Improving Transparency in Autonomous Agents. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems* (Virtual Event, New Zealand) *(AAMAS '22)*. International Foundation for Autonomous Agents and Multiagent Systems, 1083–1091.

[187] Marissa Radensky, Doug Downey, Kyle Lo, Zoran Popovic, and Daniel S Weld. 2022. Exploring the Role of Local and Global Explanations in Recommender Systems. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.

[188] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[189] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 469–481.

[190] Kanna Rajan and Alessandro Saffiotti. 2017. Towards a science of integrated AI and Robotics. , 9 pages.

[191] Juan Rebanal, Jordan Combitsis, Yuqi Tang, and Xiang'Anthony' Chen. 2021. XAlgo: A design probe of explaining Algorithms' internal states via question-answering. In *26th International Conference on Intelligent User Interfaces*. 329–339.

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

[193] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[194] Michelle Richardson, Charles Abraham, and Rod Bond. 2012. Psychological correlates of university students' academic performance: a systematic review and meta-analysis. *Psychological bulletin* 138, 2 (2012), 353.

[195] Mark O Riedl. 2019. Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies* 1, 1 (2019), 33–36.

[196] Yao Rong, Nora Castner, Efe Bozkir, and Enkelejda Kasneci. 2022. User Trust on an Explainable AI-based Medical Diagnosis Support System. *arXiv preprint arXiv:2204.12230* (2022).

[197] Yao Rong, Chao Han, Christian Hellert, Antje Loyal, and Enkelejda Kasneci. 2021. Artificial intelligence methods in in-cabin use cases: a survey. *IEEE Intelligent Transportation Systems Magazine* 14, 3 (2021), 132–145.

[198] Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. 2022. A Consistent and Efficient Evaluation Strategy for Attribution Methods. In *International Conference on Machine Learning*. PMLR, 18770–18795.

[199] Andrew Ross, Nina Chen, Elisa Zhao Hang, Elena L Glassman, and Finale Doshi-Velez. 2021. Evaluating the interpretability of generative models by interactive reconstruction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[200] Leonid Rozenblit and Frank Keil. 2002. The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive science* 26, 5 (2002), 521–562.

[201] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[202] Tiia Ruutmann and Hants Kipper. 2011. Teaching Strategies for Direct and Indirect Instruction in Teaching Engineering. In *2011 14th International Conference on Interactive Collaborative Learning*. 107–114.

[203] Hatice Sahin, Heiko Müller, Shadan Sadeghian, Debargha Dey, Andreas Löcken, Andrii Matviienko, Mark Colley, Azra Habibovic, and Philipp Wintersberger. 2021. Workshop on Prosocial Behavior in Future Mixed Traffic. In *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 167–170.

[204] Wojciech Samek and Klaus-Robert Müller. 2019. Towards explainable artificial intelligence. In *Explainable AI: interpreting, explaining and visualizing deep learning*. Springer, 5–22.

[205] Jeff Sauro and Joseph S Dumas. 2009. Comparison of three one-question, post-task usability questionnaires. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1599–1608.

[206] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 240–251.

[207] Tobias Schneider, Joana Hois, Alischa Rosenstein, Sabiha Ghellal, Dimitra Theofanou-Fülbier, and Ansgar RS Gerlicher. 2021. Explain yourself! transparency for positive ux in autonomous driving. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.

[208] Jakob Schoeffer and Niklas Kuehl. 2021. Appropriate fairness perceptions? On the effectiveness of explanations in enabling people to assess the fairness of automated decision systems. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. 153–157.

[209] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. 2022. "There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making. *arXiv preprint arXiv:2205.05758* (2022).

[210] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2017. Design and evaluation of a short version of the user experience questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence, 4 (6), 103-108.* (2017).

[211] Hendrik Schuff, Alon Jacovi, Heike Adel, Yoav Goldberg, and Ngoc Thang Vu. 2022. Human Interpretation of Saliency-based Explanation Over Text. *arXiv preprint arXiv:2201.11569* (2022).

[212] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[213] Prafull Sharma and Yingbo Li. 2019. Self-supervised contextual keyword and keyphrase retrieval with self-labelling. (2019).

[214] Jenni AM Sidey-Gibbons and Chris J Sidey-Gibbons. 2019. Machine learning in medicine: a practical introduction. *BMC medical research methodology* 19, 1 (2019), 1–18.

[215] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. 2021. Pre-registration: Why and how. *Journal of Consumer Psychology* 31, 1 (2021), 151–162.

[216] Uri Simonsohn, Leif D Nelson, and Joseph P Simmons. 2014. P-curve: a key to the file-drawer. *Journal of experimental psychology: General* 143, 2 (2014), 534.

[217] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[218] Leon Sixt, Martin Schuessler, Oana-Iuliana Popescu, Philipp Weiß, and Tim Landgraf. 2022. Do Users Benefit From Interpretable Vision? A User Study, Baseline, And Dataset. In *International Conference on Learning Representations*.

[219] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. 2020. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[220] Alison Smith-Renner, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2020. Digging into user control: perceptions of adherence and instability in transparent models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 519–530.

[221] Pravik Solanki, John Grundy, and Waqar Hussain. 2022. Operationalising ethics in artificial intelligence for healthcare: a framework for AI developers. *AI and Ethics* (2022), 1–18.

[222] Tania Sourdin. 2018. Judge v Robot?: Artificial intelligence and judicial decision-making. *University of New South Wales Law Journal, The* 41, 4 (2018), 1114–1133.

[223] Aaron Steinfeld, Odest Chadwicke Jenkins, and Brian Scassellati. 2009. The oz of wizard: simulating the human for interaction research. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. 101–108.

[224] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR, 3319–3328.

[225] Harini Suresh, Kathleen M Lewis, John Guttag, and Arvind Satyanarayan. 2022. Intuitively assessing ml model reliability through example-based explanations and editing model inputs. In *27th International Conference on Intelligent User Interfaces*. 767–781.

[226] Prasanna Tambe, Peter Cappelli, and Valery Yakubovich. 2019. Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review* 61, 4 (2019), 15–42.

[227] Erico Tjoa and Cuntai Guan. 2020. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems* 32, 11 (2020), 4793–4813.

[228] Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. 2020. Sanity checks for saliency metrics. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 6021–6029.

[229] Chun-Hua Tsai and Peter Brusilovsky. 2018. Beyond the ranked list: User-driven exploration and diversification of social recommendation. In *23rd international conference on intelligent user interfaces*. 239–250.

[230] Chun-Hua Tsai and Peter Brusilovsky. 2019. Explaining recommendations in an interactive hybrid social recommender. In *Proceedings of the 24th international conference on intelligent user interfaces*. 391–396.

[231] Chun-Hua Tsai, Yue You, Xinning Gui, Yubo Kou, and John M Carroll. 2021. Exploring and promoting diagnostic transparency and explainability in online symptom checkers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[232] Raju Vaishya, Mohd Javaid, Ibrahim Haleem Khan, and Abid Haleem. 2020. Artificial Intelligence (AI) applications for COVID-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 14, 4 (2020), 337–339.

[233] Niels Van Berkel, Jorge Goncalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M Kelly, and Vassilis Kostakos. 2019. Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–21.

[234] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[235] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence* 291 (2021), 103404.

[236] Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. 2003. User acceptance of information technology: Toward a unified view. *MIS quarterly* (2003), 425–478.

[237] Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing* 10, 3152676 (2017), 10–5555.

[238] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Transparent, explainable, and accountable AI for robotics. *Science robotics* 2, 6 (2017), eaan6080.

[239] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

[240] Frank Edward Walter, Stefano Battiston, and Frank Schweitzer. 2008. A model of a trust-based recommendation system on a social network. *Autonomous Agents and Multi-Agent Systems* 16, 1 (2008), 57–74.

[241] Clarice Wang, Kathryn Wang, Andrew Bian, Rashidul Islam, Kamrun Naher Keya, James Foulds, and Shimei Pan. 2022. Do Humans Prefer Debiased AI Algorithms? A Case Study in Career Recommendation. In *27th International Conference on Intelligent User Interfaces*. 134–147.

[242] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[243] Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*. 318–328.

[244] Yunlong Wang, Priyadarshini Venkatesh, and Brian Y Lim. 2022. Interpretable Directed Diversity: Leveraging Model Explanations for Iterative Crowd Ideation. In *CHI Conference on Human Factors in Computing Systems*. 1–28.

[245] Olivia Watkins, Sandy Huang, Julius Frost, Kush Bhatia, Eric Weiner, Pieter Abbeel, Trevor Darrell, Bryan Plummer, Kate Saenko, and Anca Dragan. 2021. Explaining robot policies. *Applied AI Letters* 2, 4 (2021), e52. https://doi.org/10.1002/ail2.52

[246] Jian Wei, Jianhua He, Kai Chen, Yi Zhou, and Zuoyin Tang. 2017. Collaborative filtering and deep learning based recommendation system for cold start items. *Expert Systems with Applications* 69 (2017), 29–39.

[247] Christina Wiethof and E Bittner. 2021. Hybrid intelligence-combining the human in the loop with the computer in the loop: a systematic literature review. In *Forty-Second International Conference on Information Systems, Austin*.

[248] Allan Wigfield and Jenna Cambria. 2010. Expectancy-value theory: Retrospective and prospective. In *The decade ahead: Theoretical perspectives on motivation and achievement*. Emerald Group Publishing Limited.

[249] Yufei Xia, Chuanzhe Liu, YuYing Li, and Nana Liu. 2017. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert systems with applications* 78 (2017), 225–241.

[250] Qian Yang, Nikola Banovic, and John Zimmerman. 2018. Mapping machine learning advances from hci research to reveal starting places for design innovation. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–11.

[251] Shuo Yang, Mohammed Korayem, Khalifeh AlJadda, Trey Grainger, and Sriraam Natarajan. 2017. Combining content-based and collaborative filtering for job recommendation system: A cost-sensitive Statistical Relational Learning approach. *Knowledge-Based Systems* 136 (2017), 37–45.

[252] Chih-Kuan Yeh, Been Kim, Sercan O Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. 2019. On completeness-aware concept-based explanations in deep neural networks. In *Advances in Neural Information Processing Systems*, Vol. 32.

[253] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[254] Asano YM., Rupprecht C., and Vedaldi A. 2020. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations*.

[255] Ofra Amir Yotam Amitai. 2022. "I Don't Think So": Summarizing Policy Disagreements for Agent Comparison *(AAAI '22)*.

[256] Wencan Zhang, Mariella Dimiccoli, and Brian Y Lim. 2022. Debiased-CAM to mitigate image perturbations with faithful visual explanations of machine learning. In *CHI Conference on Human Factors in Computing Systems*. 1–32.

[257] Wencan Zhang and Brian Y Lim. 2022. Towards Relatable Explainable AI with the Perceptual Process. In *CHI Conference on Human Factors in Computing Systems*. 1–24.

[258] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W Bruce Croft. 2018. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th acm international conference on information and knowledge management*. 177–186.

[259] Yue Zhao, Maciej K Hryniewicki, Francesca Cheng, Boyang Fu, and Xiaoyu Zhu. 2018. Employee turnover prediction with machine learning: A reliable approach. In *Proceedings of SAI intelligent systems conference*. Springer, 737–758.

A FOUNDATION OF XAI USER STUDIES

Through analyzing references in the core papers, we provide XAI researchers with several indispensable literature sources in this field, which can inspire them when organizing their projects. In total, there are over 3000 references from all the core papers, and we pay close attention to the references which are cited at least by ten core papers (ca. 50 papers). In Table 8, we categorize these papers according to their topics. The first group of papers are survey papers about XAI which are thoroughly discussed in the Section 2. For the theory of XAI, Miller [160] propose to build XAI on social sciences such as cognitive science and psychology, while Wang et al. [242] and Liao et al. [150] provide theoretical guidelines for designing XAI frameworks. An important class of references are XAI methods and the most popularly used ones are listed in "XAI Methods". Suggested by [132, 134], the explanations should be sound and complete and thus bring positive impact on users. Another motivation for XAI is that should assist users in building mental models of the AI systems [133]. Previous user studies for ML systems or for explainable interfaces that are referenced for comparisons or serve as templates of user study design. In the end, we list several general works about user trust that may go beyond the scope of XAI.

Topic	Fundamental works
Surveys of XAI	[66], [154], [1], [3], [85], [103], [19]
Theories for XAI	[160]: social sciences, [242]: theory for XAI design, [150]: a question bank for XAI design
XAI Methods	[91]: a survey, [192]: LIME, [193]: Anchors, [155]: SHAP, [119]: TCAV, [101]: explaining recommendation systems, [40]: intelligible models, [125]: influence function, [239]: counterfactual explanations, [224]: Integrated Gradient (IG), [217]: saliency maps for images, [212]: GradCAM
Principles of Explanations	[132, 134]: completeness and soundness, [133]: helping users build mental models
User studies for ML	[38]: image retrieval algorithm for medical uses, [129]: interactive model
User studies for XAI	[28]: justice perceptions, [63]: fairness [140]: human-AI team, [106]: usability, [118, 153, 168, 185, 188]: understanding, [34, 37, 45, 115]: trust and understanding
Trust	[36]: trust (calibration), [145]: trust in automation, [253]: impact of model accuracy on trust, [56, 123]: impact of system transparency on trust,

Table 8. Fundamental works of the core papers (categorized according to topics).

B MODELS AND EXPLANATIONS IN XAI USER STUDIES

The details about models and explanations used in our studied papers are given in Table 9. Black-box models are dominant in the current human-AI interaction research area as we can see that more black-box models are studied. Local feature explanations are popularly used such as LIME [192] and SHAP [155]. For recommendation systems, content-based and hybrid explanations are widely used explanations. A content-based explanation is a single-style explanation coming from a content-based recommendation system, while a hybrid explanation contains multiple explanation styles such as user-based or item-based, which is provided by a hybrid recommendation system [79, 127, 128]. For instance, Dominguez et al. [64] provide a content-based explanation as *"Painting A is 85% similar to the Painting B that you like"*. Tsai and Brusilovsky [230], however, use hybrid explanations in textual and visual explanation formats.

		White-box	Black-box	Others
Feature-based	local	[2, 40, 115]: GAMs [2, 26, 45, 243]: linear models [26]: decision tree [18]: Bag-of-Words [231]: Chatbot	[26]:Gradient Boosting Trees+LIME [26, 115]:Gradient Boosting Trees+SHAP [26, 169]: Random Forest+LIME [26, 50]: Random Forest+SHAP [96]: DNN (tabular) [7, 21, 42, 173, 218]: CNN (image or VQA) [18, 22, 23, 96, 139, 211]: DNN (nat. language) [32]: Task-specific model+SHAP [244]: Directed-Diversity+word attributions [193]: (DNN(VQA),Undisclosed)+LIME [93]: Undisclosed+(LIME+SHAP) [219]: Bayes model (nat. language) [8]: SVM/MLP+LIME [139, 140] linear SVM	
	global	[185, 243]: linear models [63]: linear model +decision boundary explanation [115]: GAMs [18]: Bag-of-Words +Linear Surrogate Model	[18]: DNN (nat.language) +Lin. Surrogate Model	
Example-based		[243]: linear models [115]: GAMs [63]: linear model [231]: Chatbot	[31, 37]: CNN (images) [218]: INN (images) [34]: Undisclosed (simulated AI) [96]: DNN (nat. language) [140]: linear SVM	[199] GAN+VAE
Counterfactual		[243]: linear models [130, 131]: Decision Tree [115]: GAMs	[209]: random forest [258]: DNN (audio) [218]: INN (images) [244]: Directed-Diversity (ideation task)	
Concept-based			[21, 84, 142, 147, 252, 256]: CNN (images) [218]: INN (images) [258]: DNN (audio)	
Others		[60, 206]: game strategy [92, 180]: decision rules	[181, 182]: RNN (Doctor AI [47]) +Ontology expl. (Dr.XAI [182]) [225]: CNN+KNN expl. [53]: semantic segmentation [34, 53, 173]: detected objects [16]: Bayesian NN+CLUE [42, 173]: confidence scores [169] Random Forest+(MAME, theirs) [193] Anchors [191]: Interactive explanations (XAlgo) [72]: Natural Language explanations [64, 136, 187]: content-based explanations for RS [128, 175, 188, 230]: hybrid explanations for RS [82]: self-designed explanations for RS [241]: debiased RS [143]: concept map + data-driven explanations	[14]: dataset explanation [48, 148, 148, 149, 159, 196, 229] [17, 75, 106, 143, 207, 220]: designed interface

Table 9. Models and explanations in core papers. Papers are categorized according to types of explanations (**column**) and types of models (**row**). For white-box models, only model name is shown; Otherwise, "model name" + "explanation name".

C MEASUREMENT DETAILS

C.1 Trust

Table 10 lists the trust measurement. Most of the works deploy questionnaires to measure user trust (self-reported), where 7-point or 5-point Likert scale is commonly used. Many works design their own questionnaires [34, 37, 64,

	Studied Paper	Metric	Definition Source	Detail
Observed	[181]	Weight of Advice (WOA)	-	Degree to which the algorithmic suggestion influences the participant's estimate.
	[140, 196, 206, 225, 243]	Agreement rate	-	Percentage of cases in which participants agree with the model. [243] defines the *appropriate trust, overtrust* and *undertrust.* [206] defines as *adherence*
Self-reported	[14, 53, 115]	Trust in Automation	[112]	On the 7-point Likert scale. [14] adapts the questions.
	[92, 136]	General trust in technology	[124]	On the 5-point Likert scale.
	[149]	Human-Computer Trust	[156]	On the 7-point Likert scale. [149] adapts the questions.
	[175]	Trust-TAM (Technology Acceptance Model)	[27]	On the 7-point Likert scale. [175] includes other self-designed questions.
	[45]	Trust in human-machine systems	[144]	On the 7-point Likert scale.
	[72]	Unified Theory of Acceptance and Use of Technology Model (UTAUT)	[236]	On the 5-point Likert scale.
	[180]	Human-Robot Collaborative Fluency Assessment	[103]	On the 7-point Likert scale
	[136]	Trusting beliefs and intentions	[158]	On the 7-point Likert scale.
	[34, 206, 209, 219, 220, 231] [37, 64, 120, 159, 196, 229]	Self-designed questionnaire	-	[37, 206, 231] are on the 7-point Likert scale. [34, 120, 159, 209, 229] are on the 5-point Likert scale. [64] rates from 0 to 100. [34, 64, 120, 159, 219, 220, 229] measure one-dimensional trust.
	[70]	Semi-structured interview	-	

Table 10. Measures of trust. The measurement is divided into two main groups: "Observed" and "self-reported" trust. The studied core papers using the same measurement are grouped together. The name and the paper reference of the used metrics are listed in the column "Metric" and "Definition Source", respectively. "-" in the column "Definition Source" means that the source is the studied paper. More details about the metrics are given in the last column.

120, 159, 206, 209, 229, 231]. To measure the trust in a objective manner, many works choose to use agreement rate of humans [140, 206, 225, 243].

C.2 Usability

Table 11 demonstrates the measures used for usability of explanations. We divide usability into five sub-categories: workload (cognitive load), helpfulness, satisfaction, debugging and ease of use and others. User perceptions of workload, helpfulness, satisfaction and ease of use are subjective and often measured with questionnaires. However, for debugging tasks, it can be measured objectively such as using accuracy of user confirming the correctness of answers from a question-answering model and the time for solving this task [120].

C.3 Understanding of Explanations

For novel or cognitively challenging types of explanations, it makes sense to verify whether users can make use of the information provided through the explanation. Usually these types of tests are conducted in combination with other measures to establish if the explanations are correctly understood by users and can thus be processed as intended.

In the domain of conceptual explanations [119, 126], such kind of understanding questions are common, to assess semantic coherence of automatically discovered concepts [84, 142, 147, 252]. Assignment tasks, where novel instances should be assingned to existing clusters are commonly used as proxy to measure the intelligibility [84, 142, 169, 252]. Another option is to assess how good the cluster can be described in natural language which is often referred to as *describability* [84, 142, 147]. Apart from conceptual explanations, Zhang and Lim [257] ask multiple choice questions to

	Studied Paper	Metric	Definition Source	Detail
Workload	[17, 53, 64, 115]	NASA TLX	[95]	
	[2]	Memory Performance	-	
Helpfulness	[34, 173, 244] [2, 256, 257]	Self-designed questionnaire	-	[34, 173, 244] are on 5-point Likert scale [2, 256, 257] are on 7-point Likert scale
	[82]	Rating	-	Rating from 1 to 5
Satisfaction	[64, 159, 219, 220, 229–231]	Self-designed questionnaire	-	[159, 229, 230] are on 5-point Likert scale [219, 220, 231] are on 7-point Likert scale [64] rates from 0 to 100
	[92, 128]	User experience of recommendation system	[124]	[92] adapts the questions on the 5-point Likert scale [128] adapts the questions on the 7-point Likert scale
	[32, 181]	Explanation Satisfaction Scale	[105]	[181] are on 5-point Likert scale [32] are on 6-point Likert scale
Debugging	[21]	Number of identified bugs	-	Questions about bug identification and solutions
	[120]	Accuracy (percentage of correct answers) and time	-	Task is to determine the correctness of model answers
	[185]	Deviation between human's and model's predictions	-	Model's predictions are buggy and human's predictions should be different.
	[218]	Accuracy (percentage of correct answers)	-	Task is to identify (ir)relevant features
Ease of use and others	[2, 17, 106, 120, 244]	Self-designed questionnaire	-	[120, 244] are on 5-point Likert scale [2, 106] are on 7-point Likert scale
	[207]	AVAM and UEQ-S	[102, 210]	Autonomous Vehicle Acceptance Model Questionnaire (AVAM) [102] User Experience Questionnaire-Short (UEQ-S) [210] Both on the 7-point Likert scale
	[199]	Single Ease Question (SEQ)	[205]	On the 7-point Likert scale
	[21]	User Engagement Scale (UES)	[178]	On the 7-point Likert scale
	[130, 131]	System Causability Scale	[107]	On the 5-point Likert scale
	[53]	System Usability Scale	[33]	On the 5-point Likert scale
	[143, 148]	semi-structured interview	-	-

Table 11. Measures of usability. The measurement is divided into five categories. The studied core papers using the same measurement are grouped together. The name and the paper reference of the used metrics are listed in the column "Metric" and "Definition Source", respectively. "-" in the column "Definition Source" means that the source is the studied paper. More details about the metrics are given in the last column.

verify if users understand the differences between the acoustical cues presented and evaluated which cue differences were most noticeable. Wang et al. [244] prompt users explicitly if the found the explanation easy to understand.

Research questions and Findings. Laina et al. [142] found that feature vectors obtained by constrastive learning approaches such as MoCo [97] or SeLa [254] allow for clusters that are almost as interpretable as human labels. Leemann et al. [147] show the similarity of ResNet-50 embeddings allows to predict how semantically coherent users find a cluster of images. For the acoustical cue, Zhang and Lim [257] found that shrillness and speaking rate where most often recognized. Wang and Yin [243] found that users reported they understood all types of explanations well without significant differences.

D FINDINGS

When using explanation types as the evaluation dimension, many works compare their effects without comparing to a control group (baseline) without explanation methods.] argue that many works have proven the usefulness of explanations and therefore no need to include such a control group. table 12 summarizes the findings of the comparison among different explanations.

		Evaluation Dimension: Explanations Effect comparison among **different explanations**
Trust		[225]: example-based explanations are positive in trust building
		[34]: deductive (rule-based) explanations > inductive (example-based) explanations in decision-making tasks, but contrary in proxy tasks
		[37]: different explanations positively affect different beliefs of trust
		[229]: proposed explanation interfaces (different visualizations), SCATTER > RANK and SCATTER > TUNER but insignificant
Fairness		[63]: sensitivity- and case-based explanations are rated as least fair when they expose a bias of the model
		[241]: acceptance of the gender-aware career recommender > gender-debiased
		[94]: significant preference for equalizing false positives over equalizing accuracy
		[209]: the amount of information positively relates with perceived fairness
		[14]: data-centric explanations that indicate balanced training data raise the fairness rating
Understanding	Obj.	[257]: Cues and Counterfactuals > Saliency (audio data)
		[2]: Sparse Lin. > COGAM > GAM
		[169]: MAME > SP-LIME
		[16]: CLUE > Sensitivity, Human CLUE, Random (for uncertainty)
		[31]: Natural images > synthetic (activation prediction)
		[218]: Counterfactuals (INN) = (proposed) Baseline Expl. > Concepts
		[193] Anchors > LIME
	Sub.	[187]: local+global explanation > local/global explanation
		[37]: example-based explanations (normative/comparative) improve the subj. understanding
		[96] LIME ≥ Composite, Prototypes and others
		[130]: closest and plausible counterfactuals, difference insignificant
		[187]: local+global explanation > local/global explanation
Usability		[2]: sLM ≤ COGAM < GAM, insignificant for self-reported cognitive load
		[32]: contextualizing/exploration improve user's satisfaction, but no significant impact when interacting both factors
		[106]: diff. expl. (e.g. local expl., counterfactuals,...)
		[115]: GAM vs. SHAP, pos. for cognitive load
		[64]: diff. interfaces, pos. for cognitive load
		[257]: counterfactual+cues > saliency, pos. for helpfulness
		[82]: DEAML > EFM (feature-level expl.) > PAV ("people also viewed" expl.) for usefulness in RS
		[173]: Salient video segments > Confidence scores, Component combinations shown for helpfulness
		[34]: deductive (rule-based) has higher cognitive load than inductive (example-based) in proxy tasks, deductive (rule-based) > inductive (example-based) in helpfulness in decision-making task
		[130]: closest and plausible counterfactuals, difference insignificant
		[128]: text explanation > visual explanations in user experience (e.g., satisfaction)
		[229]: proposed explanation interfaces (different visualizations), SCATTER > RANK and TUNER > SCATTER in satisfaction, RANK > SCATTER and TUNER > SCATTER in usefulness, but all insignificant
		[218]: Counterfactuals (INN) = (proposed) Baseline Expl. > Concepts in bias detection
Human-AI Team Performance		[34]: both deductive (rule-based) explanations and inductive (example-based) explanations are positive, no significant difference

Table 12. User study findings when using model **explanations** as evaluation dimensions and comparing different explanation types on measured quantities.