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

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

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

Index Terms—XAI, Human-centered XAI, Explainable ML, User Study, Human-AI Interaction

1 Introduction

Artificial Intelligence (AI) is driving digital transformation and is already an integral part of various everyday technologies. Recent developments in AI are essential to progress in fields such as recommendation systems [98, 99, 100], autonomous driving [101, 102, 103] or robotics [104, 105, 106]. Moreover, AI's success story has not excluded high-stakes decision-making tasks like medical diagnosis [107, 108, 109], credit scoring [110, 111, 112], jurisprudence [113, 114] or recruiting and hiring decisions [115, 116], However, the behavior and decision-making processes of modern AI systems are often not understandable, so they are frequently considered black boxes. Deploying such black-box models presents a serious dilemma in certain safety-critical domains, for instance, public health or finance [117]. This is due to the necessity for a transparent and trustworthy AI system, which is required by both practitioners (to gain better insights into system functioning) and end users (to rely on model decisions).

Methods to increase the interpretability and transparency of an AI system are developed in the research area of Explainable AI (XAI). Specifically, human-centered XAI, which addresses the importance of human stack-holders to the AI systems, has been proposed and discussed since [118, 119]. While a huge number of model explanations are available, the question of how to transparently evaluate their quality is still an open research question, and hence, extensively studied in recent years. A popular taxonomy of evaluation strategies for XAI methods proposes three categories: functionally-grounded evaluation, application-grounded evaluation, and human-grounded evaluation [120]. 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 [121] for review), however, the difficult comparability between different automatic evaluation measures is a common problem [122, 123]. Another drawback of automated measures is that there is no guarantee that they truly reflect humans' preferences [40, 124]. Consequently, user studies in XAI, especially when moving towards real-world products, are inevitable if one wishes to test more general beliefs of the quality of explanations [16]. However, only a small portion (about 20 %) of XAI evaluation projects consider human subjects [121]. There exist efforts in developing taxonomies or introducing the definitions or implications of different human-centric evaluations [125, 126, 127], but the recent generation of user studies and their findings have not been systematically discussed yet. Moreover, Yang et al. [128] point out that XAI is growing separately and treated differently in different communities (e.g., machine learning and HCI). Hence, effective guidance in XAI user study design is crucial to better let both XAI algorithm 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 machine learning (ML)/AI and

- *Y. Rong, Tobias Leemann, Thai-trang Nguyen, Lisa Fiedler and Gjergji Kasneci are with the Department of Computer Science, University of Tübingen, 72076, Germany. E-mail: {yao.rong, tobias.leemann, gjergji.kasneci}@uni-tuebingen.de, {thai-trang.nguyen, lisa.fiedler}@student.uni-tuebingen.de*
- *Peizhu Qian and Vaibhav Unhelkar are with the Department of Computer Science, Rice University, 77005, Houston, USA. E-mail: {pq3, vaibhav.unhelkar}@rice.edu*
- *Tina Seidel and Enkelejda Kasneci are with the Department of Department of Educational Sciences, Technical University of Munich, 80335, Germany. E-mail: {tina.seidel, enkelejda.kasneci}@tum.de*

Corresponding author: yao.rong@uni-tuebingen.de

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

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.

human-computer interaction as a prerequisite, the scope of this paper covers the main venues in these fields. We identified a total of 97 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:

1) To offer an overview of the foundational work of user studies in XAI, we investigated references of all 97 core papers in a data-driven manner. Likewise, we analyzed follow-up works building on these core papers (identified through citations of core papers) to reveal the fields impacted by XAI user evaluations (Section 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 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 a summary of the state-of-the-art.
4) Based on the examined user study details and their best-practice findings, we synthesize guidelines for designing an effective user study for XAI (Section 6).
5) Beyond the user study design, we discuss potential paradigms of AI systems understanding humans in the context of e.g., theory of minds, as well as other future research directions (Section 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.

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 [129, 130, 131, 132, 133, 134], the authors aim to cover many facets of XAI technologies ranging from problem definitions, goals, AI/ML model explanations to evaluation measures, while in [135] 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, [136, 137, 138] investigate explanations for Deep Neural Networks (DNNs), where models often take images as input [136, 137]. Joshi et al. [138], 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. [139] provide a literature review for causal explanations. A systematic literature review on explanations for advice-giving systems is conducted in [140]. Among these surveys focusing on general XAI technologies, evaluation measures are only briefly examined.

One challenge in XAI research is to evaluate and compare different explanation methods, due to the multidisciplinary concepts in interpretability/explainability [120, 121, 141]. 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 [120, 121]. Many researchers seek solutions to evaluate explanations automatically. A comprehensive literature review with a focus on these functionally-grounded evaluation methods (without human subjects) can be found in [121]. 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 [120, 142].

For instance, Chromik and Schuessler [126] propose a taxonomy on XAI evaluations involving humans. Mohseni et al. [127] 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 [125] places more focus on psychometric evaluations by proposing a conceptual model of the XAI process and specifying four key components that should be evaluated: explanation goodness and satisfaction, (user's) mental models, curiosity, trust and performance. Beyond assessing evaluation methods, XAI applications are designed to eventually support decision-making and benefit end users. A recent review by Lai et al. [143] considers studies on collaborative Human-AI decision-making, which may include AI agents providing explanations. Success in human-AI decision-making tasks can be seen as one amongst many other ways to evaluate the effect of explanations. Ferreira and Monteiro [144] present a review of the user experience of XAI applications to answer who uses XAI, why, and in which context (what + when) the explanation is presented.

Closer to our focus on user studies concerning XAI, Liao

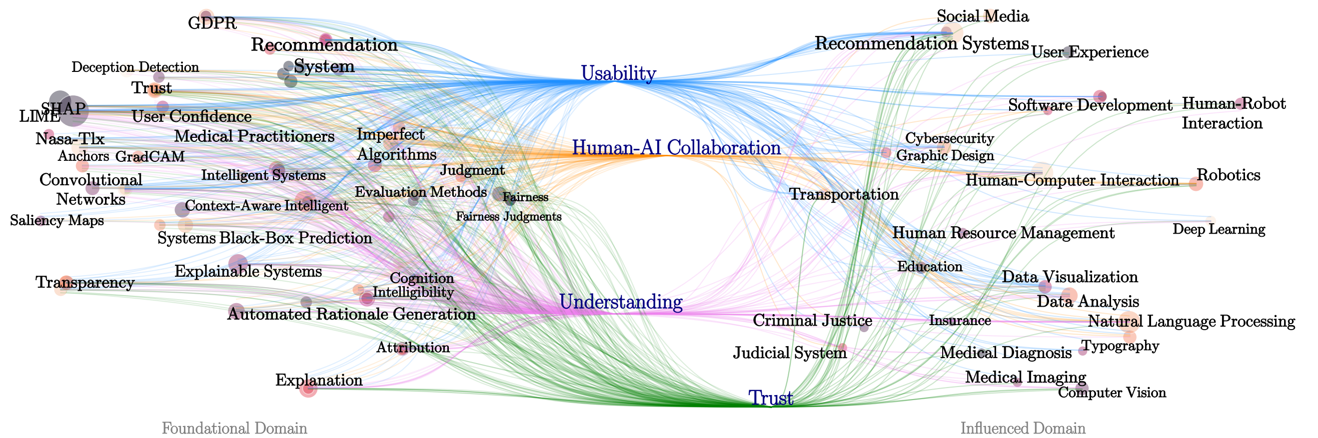

Fig. 1: Illustration of the **foundational** research domains (**Left**): Each dot represents a referenced paper, whose size reflects the number of studied core papers referring to it. Illustration of **influenced** research domains (**Right**): Each dot represents a research topic, whose size refers to the number of papers on the same topic. For a clear depiction, only several important research domains are labeled with text. Each line represents a reference link. Core paper categories are in blue (**Middle**). Lower transparency of lines indicates more links between core papers and that research domain.

et al. [142] study user experiences with XAI to reveal pitfalls of existing XAI methods, underscoring the important role of humans in XAI development. As suggested by Doshi-Velez and Kim [120], a human-subject experiment needs to be designed sophisticatedly to reduce confounding factors. In contrast to previous surveys on XAI, we aim to provide XAI researchers and practitioners with a comprehensive overview of the research questions explored in user studies, along with thorough information on experimental design. To this end, we present a practical guideline in user study design, which can be used as a starting point for future exploration of human-centric XAI applications.

3 Foundations and Impact of 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 [135].

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.

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

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

Specifically, we included the recent *five* years of CHI, IUI, UIST, CSCW, FA(cc)T, ICML, ICRL, NeurIPS, and AAAI. As we aim at analyzing human user evaluation of advanced model explanations, we ran search queries involving keywords from the two groups "explainable AI" and "user study", as listed in the Table 2. We selected the papers containing at least one keyword from each group, resulting in over one hundred papers. Then, we thoroughly studied these papers and filtered out papers that did not fulfill the criteria: (1) deploying explainable models or techniques and (2) conducting an assessment with human subjects.

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 [120], *interpretability* in the context of ML systems is defined as the ability to explain or present model predictions in understandable terms to a human. Beyond fostering comprehension, the authors argue that interpretability can assist in qualitatively ascertaining whether other desiderata, such as *usability* and *trust* are met. During a profound study of the relevant literature that was previously selected, we identified four sensible categories, that are derived from the considered dependent variables in user studies (desiderata of interpretability). These four categories are **trust, understanding, usability**, and **human-AI collaboration performance**. In Table 1, the studied papers are categorized according to the measured quantities.

In the context of interpretability, trust is when a user has confidence in and relies on an ML model's predictive quality and confidently lets the model decide. If its errors align with expert errors, the user might still trust the model, but if the model errs where the experts succeed, trust diminishes [141]. Understanding, in the context of interacting with an ML model, refers to a user's grasp or "mental model" of how the model operates, and this knowledge grows from using the system and from clear explanations about it [142]. According to [145], usability is the extent to which users can utilize a product to successfully, efficiently, and satisfactorily accomplish their intended objectives. Thus, this category encompasses user studies that employ model explanations to support users in achieving specific tasks. In usability, different aspects are measured, for instance, whether the system is easy to use or how much cognitive load it requires. The aspect "undesired behavior detection" relates to use cases where explanations uncover model discriminatory behaviors, such as the utilization of undesired features. Human-AI collaboration performance is related to scenarios where the AI system provides its predictions, but humans retain the final decisions [89]. In this case, model explanations are deployed to reach a performance superior to that of the AI system or the human decision-maker alone. These categories cover different dependent variables of interest in the reviewed user studies, primarily related to how XAI methods function. These functions mainly tie to the models' reasoning and knowledge representation. A wider perspective on XAI, which assesses generalization or robustness, remains an important field for future exploration through user studies.

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" [146]). The same API was used to extract abstracts from Google Scholar for all references and citations. Based on the paper titles and abstracts, we utilized GPT-4 [147] to tag the papers with keywords and subsequently reviewed the sensibility of these keywords manually. We visualized papers in a 2-dimensional semantic space according to their keyword embeddings using t-SNE [148].

3.2 Foundations of User Studies

We illustrate the research domains that are fundamental to XAI user studies in Figure 1 (**Left**). Note that for presentation clarity, we only visualized works that were used as references in at least five of the core papers. We can see that model explanations and interpretability are essential components, which include papers introducing explanation methods, such as LIME [149], SHAP [150] or Anchors [55]. Convolutional networks, commonly employed in experiments, utilize tools like GradCAM [151] or other saliency maps to generate model explanations. Notably, many research papers appear within the domain of recommender systems, because many XAI user studies are conducted in the context of recommendation solutions. The EU's General Data Protection Regulation (GDPR) [152] is mentioned by many core papers due to the ongoing debate on the "right to explanation" [153], 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 are underrepresented. For instance, only a small dot labeled "Cognition" can be observed in Figure 1, which is a paper that proposes to enhance XAI theory with social sciences such as cognitive science and psychology [18]. In combination with the lack of references in psychology, this indicates that only a few XAI user studies attempt to evaluate XAI from human psychological aspects. We highlight a nascent research domain of XAI frameworks based on human cognition and behavior theories [142]. 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 1 (**Right**) demonstrates the "consumers" of the human-centered XAI core papers (i.e., research domains 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 noticed that trust measurement user studies impact a variety of applications, particularly recommendation systems. Likewise, studies on user understanding span a wide range of applications. Papers on usability have a significant impact on fields like graphic design and education. In these areas, models frequently serve as tools to ease the burden on end users. Human-AI collaboration measures particularly promote the further development of robotics and or natural language processing. The 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.

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 Comprehensive User Study Analysis

In this section, we present details of the covered XAI user studies. We first introduce some commonly used AI models and explanation techniques (Section 4.1), followed by a discussion of application domains and measures with respect to the four measured quantities. The experimental designs, as well as analysis tools are presented in Section 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

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

		White-box	Black-box	Other
Feature-based	**local**	[21, 48, 156] [12, 22, 39] [6, 50]	[21, 45, 49, 55] [29, 34, 72, 92] [35, 39, 40] [42, 47, 65] [54, 57, 58] [50, 56, 71] [40, 41, 89] [25, 59, 90] [43, 60, 95]	
	global	[12, 53, 74] [21, 50]	[50]	
Example-based		[12, 21, 43] [6, 74, 96]	[17, 52, 57] [13, 25, 40]	[32] (generative models)
Counterfactual		[12, 37] [21, 82]	[27, 100] [57, 65]	
Concept-based			[61, 62, 71] [63, 64, 67] [57, 100]	
Other		[11, 88] [7, 10]	[1, 9, 15, 157] [3, 13, 51] [3, 56, 58] [36, 49, 55] [16, 28, 85] [33, 38, 68]* [8, 23, 76] * [69, 70]*	[2] [18, 19, 20] † [20, 26, 84] † [66, 81, 83] † [14, 30, 85] † [5, 91] †

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.

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 [154]. First, there are intrinsically interpretable models, also known as *white-box models*. For instance, white-box models include decision trees and linear models. Second, there are *black-box models* that provide no parameter access or are too complex to be explained in a human-understandable way [155]. These include ensembling techniques such as Random Forests or neural models.

As for explanation techniques, we identified five key types in the scope of the surveyed papers (rows of Table 3). Most frequently used are feature-based (attribution) explanations, for instance, SHAP (Shapley additive explanations [150]) and LIME (Local Interpretable Model-Agnostic Explanations [149]). There is a clear differentiation between local, instance-wise, explanations and global explanations that apply to the model in its entirety. For instance, the weights of a linear model have a global scope. This differentiation is common among these feature-based explanations, where most of the papers using local explanations. Other popular explanation types are example-based explanations, counterfactual explanations, which aim at providing actionable suggestions for attaining a user-preferred prediction by changing certain input features, and concept-based explanations, which use meaningful high-level concepts such as objects or shapes to explain a prediction.

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

4.2 Measurements

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

4.2.1 Trust

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

Trust measures used in much of the existing research can be divided into two groups: *self-reported* and *observed* trust [158]. 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 3 in Appendix, trust measures in these two groups are listed. The agreement rate of users with the model decisions is commonly used [9, 11, 12, 25] as a measure of observed trust. Parallel to observed trust measurement, van der Waa et al. [159] ascribe the user's alignment behaviors to the *persuasive power* of model explanations, i.e., the capacity to convince users to follow model decisions despite the correctness. As an extension, trust calibration is defined based on this measure. For example, a high agreement rate to wrongly made decisions represents *overtrust*, while a low agreement rate to correct decisions means *undertrust* [12]. In self-reported measurements, researchers either utilize well-developed questionnaires or self-designed ones, with the exception of [4] which conducts a semi-structured interview to explore user opinions. Several works [6, 11, 13, 16, 17, 18, 19, 24, 27] propose their own questionnaires. Among these, a subgroup [13, 16, 18, 19, 24] simply asks users to rate a single statement such as "I trust the system's recommendation/decision", which is named as one-dimensional trust by [8]. When deploying previously proposed questionnaires [2, 3, 5, 7, 8, 10, 21, 22, 23, 160], Trust in Automation [161] is the most commonly used one, in which the underlying constructs of trust between human and computerized systems are explored.

4.2.2 Understanding

An important goal of explanation techniques is to foster users' understanding of complex ML systems. An important separation has to be made between users' perceived understanding and their actual comprehension of the underlying model, as the two often do not agree [35, 40]. Cheng et al. [22] explicitly differentiate between *objective* understanding and self-reported understanding, which we term *subjective* understanding in this work. While subjective understanding is usually measured through questionnaires, measuring objective understanding requires a proxy task where the users' understanding is put to a test. Additionally, user studies can be run to assess how well users can understand the explanation itself (and not the underlying model).

This can be an important sanity check and is particularly used in the domain of conceptual explanations [62, 162], where the intelligibility of concepts needs to be verified. We refer to the third category as *understanding of explanations* but defer its detailed findings to Appendix C.3.

Objective Understanding. Works in the subdomain of objective understanding deploy proxy tasks to verify users' understanding of a model's inner workings. The most commonly considered domain in works on understanding is finance [35, 39, 40, 47, 48, 49, 53] followed by image classification [13, 21, 52]. One of the most critical design choices when assessing objective understanding is the selection of a suitable proxy task. Doshi-Velez and Kim [120] argue that the task should *"maintain the essence of the target application"* that is anticipated. One of the most prominent tasks is forward simulation [120, 141]. 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* [141]. However, scholars have designed many more tasks to quantify understanding and applied them across a variety of data modalities (cf. Table 2 in Appendix 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* [120] 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 [141] pointed out that simulatability can only be a reasonable measure, if the model is simple enough to be captured by humans and that simpler tasks are required otherwise. An example could be a *feature importance* query, where users have to tell which features are actually used by the model. A directed and more local version of this task is *marginal effects queries*, where the subjects predict how changes in a given input feature will affect the prediction (e.g., *"Does increasing feature X lead to a higher prediction of Y being class 1?"*). Because explanations should allow the identification of weaknesses in models, the task of *failure prediction* measures the accuracy of users' prediction when the model prediction is wrong.

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

Most of the works assess the subjective understanding of a user with a post-task questionnaire. Guo et al. [7] adapted a popular questionnaire designed for recommendation systems by Knijnenburg et al. [163], while Bell et al. [39] accommodated the questionnaire which originally intended to measure the intelligibility of differenet explanations by Lim and Dey [164]. On the other hand, agreement to simple subjective statements such as *"I understand this decision algorithm"* [22], *"I understand how the AI..."* [13, 17] or *"The explanation(s) help me to understand..."* [33] can be collected to assess subjective understanding.

4.2.3 *Usability*

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

Based on the measurements in the user studies, we refined usability into measures of helpfulness, workload (cognitive load), satisfaction, ease of use and detecting undesired behaviors of the system, as shown in Table 1. To assess workload (cognitive load), NASA-TLX scale [165] is used in [3, 6, 16, 21, 66], while Abdul et al. [48] measure cognitive load by capturing the log-reading time of memorizing the explanation. Most of the works use self-designed questionnaires or statements to measure satisfaction [6, 16, 18, 19, 29, 30, 69, 70], however, the Explanation Satisfaction Scale [166] can be deployed as an established alternative [1, 47]. Helpfulness can be assessed by simply asking for subjective ratings of the explanations for accomplishing a specific task [13, 46, 56, 65, 67, 68]. Colley et al. [3] use an adapted version of the System Usability Scale proposed in [167].

Two primary concerns lead to undesired behaviors: bugs and inherent biases in the system. Model explanations should illuminate these issues for users. In debugging applications, the effectiveness of explanations is evaluated by objective performance measures, such as the number of bugs identified [71], the share of participants that identify a certain bias [57, First Experiment] or by the deviations between model predictions and human predictions for unusual samples [53]. An exhaustive overview of measures for usability is given in Table 4 of the Appendix. The perception of users regarding fair treatment by a system has primarily been researched in high-stakes applications such as granting loans [27] or granting bail for criminal offenders [73, 74, 75]. For example, [73, 74, 75] investigate the fairness of COMPAS, a commercial criminal risk estimation tool that was used in the US to help make judicial bail decisions. It is also considered in everyday use-cases such as news [38] and music [77] recommendations, or possible career suggestions [76], where a bias in the underlying system can be to the detriment of the user. As the assessment of fairness is a very subjective matter, questions regarding perceived fairness are prevalent, e.g., "how the software made the prediction was fair" [74], which can be answered on 5- or 7-point Likert scales [2, 27, 38, 73, 74, 75]. Among these works, an effective explanation is the one that can increase human perceptions of fairness, except in [74], where the authors find that models which provide 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 [74].

4.2.4 *Human-AI Collaboration Performance*

The goal of human-AI teaming is to improve the performance in AI-supported decision-making above the bar set by humans or an AI alone [89]. Improving human performance with the help of AI has been considered in

	Experimental Design		
	Between-Subjects	Within-Subjects	Mixed
Papers	[5, 7, 8, 12, 15, 27, 59]		
	[17, 21, 22, 23, 25, 72]	[1, 3, 4, 9, 19]	
	[11, 28, 32, 40, 46, 95]	[10, 18, 21, 24, 70]	[2, 13, 16, 52, 66]
	[47, 49, 50, 51, 53, 84]	[13, 26, 35, 52, 91]	[10, 28, 34, 64, 74]
	[36, 37, 38, 39, 43, 56]	[57, 71, 78, 81, 93]	[12, 33, 65, 83, 149]
	[29, 30, 54, 90, 92, 96]	[6, 62, 63, 67, 69]	[61]
	[12, 38, 57, 58, 73]	[14, 41, 45, 60, 68]	
	[75, 76, 77, 78, 82]		

TABLE 4: Experimental designs in core papers.

games [10, 88], question answering tasks [89, 91], deception detection [25, 90] and topic modeling [29, 30].

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

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 4 and Figure 2, respectively.

4.3.1 Between-subjects

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

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

The statistical analysis directly follows from this design. If one is interested in identifying significant differences between the groups, common statistical hypotheses tests are used. For overall comparison, one or two-way ANOVA tests are the most commonly used statistical tool. Interesting post-hoc comparisons between two groups can be made with a standard T-test, if the data is normally

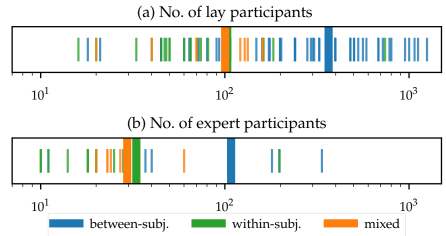

Fig. 2: Distribution of participant numbers in the surveyed user studies by design and participant type (each bar represents one study). Per-design means are indicated in bold.

distributed with equal variance, or by using non-parametric tests such as the Wilcoxon rank-sum test (also known as Mann-Whitney U-test) for comparison of two populations (e.g, [57]) or the Tukey HSD test (e.g., [49]) for multiple populations. When running multiple post-hoc tests, some works make use of the Bonferroni correction (e.g, [57]).

4.3.2 Within-subjects

Around $30\,\%$ of the papers use the within-subjects design, where each participant sequentially passes through all conditions and provides feedback. Fewer participants are recruited in within-subjects experiments compared to the between-subjects ones. Hence, they are particularly popular when participants with restrictive characteristics, such as domain-specific professional expertise, are required. For example, Suresh et al. [9] and Rong et al. [26] recruit fourteen medical professionals and five radiologists in their user studies, respectively. The small number of medical experts contributing to the user study is a limitation [26], however, it is often the case in expert user research. Gegenfurtner et al. [168] evaluate 73 sources and point out that the majority of these studies include only five, maybe ten experts. Besides the medical domain, other works [3, 4, 19, 21] also invite subjects with particular professions such as engineers in a technology company. When no specific knowledge is required, however, participant numbers reach up to 740 also for within-subjects designs [93].

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

4.3.3 Mixed

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

5 FINDINGS OF USER STUDIES

In this section, we summarize the primary findings from the core papers. Table 5 lists findings with respect to four measured quantities. To build an overview of the findings, we divide papers according to their evaluation dimensions, i.e., the independent variables in the user studies. When using the presence of explanations as the evaluation aspect, the findings are summarized in Table 5. The listed impacts using explanations are to be seen in comparison with a control group without explanations. Effects are divided into two groups: (1) Positive effects, for example, increasing user trust or understanding; (2) Non-positive effects: the effect can be negative, or not significantly positive (neural), or a mixture of different effects (e.g., feature-based explanations have positive effects but counterfactual explanations do not). Beyond the explanations themselves, other possible evaluation dimensions such as that might have an impact on the perception of XAI, for instance, AI technology literacy, model performance, or the dimensionality of the data. Instead of using the mere presence of explanations, many works compare different explanation techniques with each other (see Appendix D for more details).

As various research questions and findings are addressed in 97 core papers, many papers compare explanation types in order to choose a preferable one, it is not possible to cover all results in one table. Based on them, we outline some interesting trends in the effectiveness of explanations on user experience: (1) Explanations are effective in improving users' subjective understanding; (2) The effectiveness of explanations in increasing user trust and usability of models is not clear; (3) Explanations are not good at convincing users that models are fair; (4) Interactivity of the model has positive impact on user trust, understanding and model usability. The first three statements can validated through the number of papers obtaining positive or non-positive effects in each category, while the last finding is extracted from Table 5 in the Appendix, which details findings with on other independent variables. We encourage the reader to consider the short summary of *primary* findings in the tables and check for further details according to their specific interests. In the following section, we highlight some findings for each category of measurement.

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

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

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

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

In comparison with the objective understanding, the research question in the subdomain subjective understanding is to find out how explanations impact user's *perceived* understanding [7, 12, 17, 22, 32, 33, 34, 37, 56]. There exist a trend of using model explanations to improve subjective understanding [13, 16, 17, 28, 34, 38, 170]. However, Chromik et

		Evaluation Dimension: Explanations Effect of explanations compared to **no explanations**	
		Positive	Non-positive / Mixed
Trust		[13]: example-based, rule-based explanations [16]: example-based explanations for recommendations [27]: feature importance [10]: decision-tree explanation for policy [28]: explanation corpus given by researchers [25]: feature-based (saliency map), example-based explanations [8]: explanations for recommendations [6]: rationale-based, example-based and feature-based (best) explanations for online symptom checkers [15]: confidence scores	[3]: positive in simulation but no improvement in real-word [1]: explanations for medical suggestions (Doctor XAI [157]) pos.for observed trust but insignificant for reported trust [12]: feature-based explanations increase appropriate trust slightly but counterfactual explanations inconclusively [21, 22, 24]: feature-based explanation, insignificant [11]: rule-based explanation, insignificant [15]: Shapley values, insignificant [29]: feature-based explanation, negative
Understanding	Obj.	[22, 53] white-box model [40] feature importance, LIME (tabular) [46] counterfactuals+cues (audio) [50] manipulatability improved by white-box log. reg. [54] saliency maps (image) [59] saliency maps for bias detection and strategy identification [12] counterfactuals+feature importance	[39]: SHAP, negative for black-box model (education domain) [39]: Insignificant difference btw. black-box and white-box models [40]: Prototypes, Anchors, LIME on textual data insignificant [46]: Counterfactuals and Concepts insignificant (audio data) [50]: Simulatability results insignificant for LIME, IG, surrogate model on BERT and Logistic Regression Model, Manipulatability insignificant for BERT [58]: Insignificant results for GRAD-CAM, Saliency Map, uncertainty scores in VQA [59] saliency maps for failure prediction (image) [60]: saliency maps, negative for a mix of three interpretation techniques in simulation task
	Sub.	[13]: example-based, rule-based explanations [28]: explanation corpus given by researchers [12]: feature-, example- and counterfactual-based [38]: explanations provided by [169] for Facebook News Feed [16]: example- and feature-based explanations [17]: example-based explanations [34]: feature importance, SHAP and LIME [35]: feature importance, SHAP	[22]: white-box model, insignificant [39]: white-box < black-box, both insignificant [31]: feature importance explanation (transparent system) can be distracting
Usability		[81]: counterfactuals, pos. for usability [16, 47]: example-based explanations, pos. for satisfaction [67]: CAM-related explanations, pos. for helpfulness [6]: rational-, feature-, example-based explanations, pos. for satisfaction [70]: content-based explanations, pos. for satisfaction [83]: explanations regarding driving information, pos. for ease of use [13]: example-based and rule-based explanations, pos. for helpfulness [71]: local, global, visual (saliency map) explanations, pos. for bug identification [65]: attribution methods and conceptual explanations, pos. for usefulness [84]: feature-based, pos. for reliability [24]: (proposed) template-based expl. pos. for debugging and usefulness [27]: feature importance, counterfactual explanations pos. for perceived fairness	[82]: counterfactuals, significant for helpfulness/usability but insignificant for usefulness [1]: ontology-based explanation, insignificant for satisfaction [65]: attribution methods and conceptual explanations, insignificant for ease of use [24]: visual explanations increases usefulness, but improvement is insignificant [3]: pos. for cognitive load/usability (simulation), but insignificant in real-world [29]: feature-based explanations, negative for satisfaction [38]: informing users about the algorithmic decisions, negative ranking scores of recommendations, insignificant for perceived fairness [27]: highlight features only, insignificant for perceived fairness [78]: insignificant in between-subjects but significant in within-subjects for perceived fairness
Human-AI Collaboration Performance		[88]: textual explanations with domain knowledge (in chess) [25, 90]: feature-based explanations [91]: exampled-based for experts, feature-based for novices [93]: contrastive explanations [13]: example-based and rule-based explanations [95, 96]: example-based explanations, attributions (AI correctness prediction) [97]: important parts in images as explanations	[25]: exampled-based, insignificant [15, 89]: feature-based explanations, insignificant

TABLE 5: 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.

al. [35] challenge the improvement in perceived understanding with the cognitive bias named *illusion of explanatory depth* (IOED) [171], which means that laypeople often have overconfidence bias in their understanding of complex systems. Their results confirm the IOED issue in XAI, i.e., questioning users' understanding by asking them to apply their understanding in practice consistently reduces their subjective understanding. Explanations can have different impacts on subjective and objective understandings [22], where white-box explanations increase objective understanding but do not have significant impact on subjective understanding. Similar disagreements have been observed in multiple other works [40, 170]. Radensky et al. [33] examine the joint effects of local and global explanations in a recommendation system and their results provide evidence that both are better than either alone.

Usability. Similar to trust, it is not clear whether explanations are effective in improving users' perceptions of helpfulness, satisfaction or other dimensions of usability. For instance, in [16, 30, 47], the explanations have a positive effect on satisfaction, while no significant effects on satisfaction are observed in [18, 19, 29, 69]. Parallel to trust, Smith-Renner et al. [29] provide evidence for the hypothesis that it is harmful to user trust and satisfaction to show explanations by highlighting the important words in a text classification task. A strong correlation between self-reported trust and satisfaction can also be observed in [3], where explanations have a positive impact in a simulated driving environment, but no significant effects when using real-world data. Beyond explanations, Nourani et al. [56] study the order of observing system weakness and strengths, which reveals that encountering weakness first results in a lower rate of usage of system explanations than encountering strength first. Schoeffer et al. [27] find out that showing feature importance scores or counterfactual explanations (or a combination of both) for explaining decisions helps increase the perceived fairness, whereas highlighting important features without scores does not. However, several studies don't show a significant difference between scenarios with and without explanations [27, 38, 78]. Effects of explanations may be dependent on input samples, as shown in [67]. The authors show that both Debiased-CAM and Biased-CAM improve the helpfulness for a weakly blurred image, however, there is no significant improvement for unblurred or strongly blurred images.

Human-AI Collaboration Performance. A strain of works [25, 88, 90, 91, 95, 96, 97] show that viewing explanations can improve human accuracy in making decisions, especially with feature-based explanations taking text data as input [25, 90, 91]. When using example-based explanations in text classification, there is no improvement in human performance [25]. Likewise, utilizing explanations has no significant impact on human performance in [89, 92], but simply showing model predictions has a positive effect in [92]. Experts and novices perceive explanations differently, for example, Feng and Boyd-Graber [91] conclude that the performance gain of novices and experts comes from different explanation sources. Paleja et al. [10] reveal that explanations can improve novices' performance but decrease experts' performance. Additionally, less complex models with explanations can better convince humans in correct decisions [90].

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 guidelines in the order of before, during and after user studies, which reflects user study design, execution and data analysis, respectively.

Before the User Study. When designing a user study, the first step is to decide what to measure. To define the measured quantities, one can consider two alternatives: using a general definition or an application-based quantity that is specific to the application at hand. The former one refers to a quantity that is borrowed from previous well-established research, such as using "trust in automation" [2, 3, 21] or "general trust in technology" [7, 23]. To further construct "trust" as a quantitative measurement, one needs to examine how existing work has conceptualized "trust" in both social sciences context as well as XAI and technical context [172]. The application-based quantity depends on the application goal, for instance in a chess game [88], the measurement is the human winning percentage with the help of model explanations (Human-AI collaboration).

From Table 5, we can see that previous works have frequently struggled to prove the effectiveness of XAI even with respect to a control group that is without explanation. When only different explanation techniques are considered, there will always be one winner explanation, but the overall benefit will remain undisclosed (see examples in Appendix D). Therefore, it is important to compare with a baseline without explanations to rigorously show the strength of XAI. When a comparative design is explicitly desired, baselines such as random explanations [28, 41, 62]).

When deploying a proxy task, its difficulty should be gauged and monitored carefully. In the past, the forward simulation task has been criticized as being unrealistically complex for domains such as computer vision [54]. Thus, other proxy tasks such as feature importance queries [57] or manipulatability checks [32, 50] were proposed. Another important point is to choose a proxy task that is simplified, but features many characteristics of the application in mind [120]. Notably, the proxy task should be designed close to the final anticipated application, as even slight differences in the tasks may void the validity of the findings on the proxy tasks in the real world [13].

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

Besides the measures introduced in Section 4.2, there are several psychological constructs that can be deployed to evaluate multiple facets of the interaction between humans and XAI. For instance, the *subjective task value* in the expectancy-value framework is often used to analyze subjective motivation to take any actions [173], which is not thoroughly studied in the XAI experience yet. The subjective task value consists of intrinsic value (enjoyment), attainment value (importance for one's self), utility value (usefulness), and cost (the amount of effort or time needed) [173, 174]. 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 [165, 175]. 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 [177]. 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 [178] and thus strengthen the credibility of a study. Expert interviews and pre-studies following a think-aloud protocol [179], e.g., in the references [32, 46], are often mentioned as helpful tools to develop the explanation system and the study design and gain first qualitative insights or complement the qualitative analysis [13, 65].

When preparing for a user study, it is important to plan for explicit steps and to have a backup plan for different situations. Before participants arrive, it is helpful to provide them with information such as where the researchers will meet with them, what they need to bring, and how they can prepare for the study. If conducting the experiment in

2. https://aspredicted.org

Before the User Study

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

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

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

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

During the User Study

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

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

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

After the User Study

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

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

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

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

person, send participants a reminder the day before and provide them with your contact in case they cannot find the experiment site or they need to cancel the experiment session. Once participants arrive, make sure the researchers have a plan that covers all stages of the experiment. The protocol should cover small details (e.g., where participants should leave their backpacks, water bottles, and lunch boxes) and plans for unexpected situations (e.g., uncooperative participants and multifunctional systems). How to obtain participants' consent should be an important part of the procedure. Additional procedure is required for obtaining consent when working with vulnerable populations (e.g., children and pregnant women), in which case alternative consent procedures might take place. Another benefit of predesigning 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 [172]. To ensure a sound experiment procedure and to protect the integrity of the data, it is worthwhile to put in much effort to design a detailed experiment script.

During the User Study. A sufficient number of participants is the prerequisite of a solid user study analysis. To get a rough estimate of common sample sizes, we refer the reader to the participant statistics in Figure 2 where we analyze the subject numbers in different experimental designs. For instance, around 350 users without any specific expertise are averagely recruited in between-subject experiments. However, we would like to underline that the required number of participants is highly specific to the study design and should be determined individually, for instance by conducting a statistical power analysis [180]. Additionally, recruited participants should have the same knowledge background as the end users that applications are designed for. For instance, when evaluating an interface explaining loan approval decisions to bank customers, it is not proper to include only students whose major is computer science, since they may have prior knowledge of how model explanations work. Note that the design of an AI application requires different audiences across the project cycle, thus model explanations need to evolve as well [181].

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. [182] justify the use of these checks without compromising scale validity. In within-subject experiments, a random order of conditions is necessary to avoid order effect [1]. Participants can learn knowledge of data or examples shown in the previous conditions, and Tsai et al. [6] choose to use a Latin square design to avoid the learning effect.

After the User Study. After the data collection, statistical tests are run to find significant effects. The applicable tests used are determined by experimental designs and the form and distribution of the data. Generally, ANOVA tests and T-test are usually used when comparing distributions between different conditions. Structural Equation Models (SEM) or multi-level models are used for mediation analysis. More details of statistic tools can be found in Section 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 [183] can be used to assess agreement beyond chance between these raters and serve as an indication for the reliability of the ratings.

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

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 the development of XAI solutions that better respond to user needs. As discussed in [118], there are two aspects of human-centered AI: (1) AI systems that understand humans with a sociocultural background and (2) AI systems that help humans understand them. The former point can guide the design of AI systems. In this section, we discuss XAI research that leverages this insight.

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

Expectancy-value Motivation Theory. Human interaction with XAI interfaces can be viewed as an activity where humans learn about the model’s inner workings through explanations and then achieve an understanding of the models. The question of how to enhance the efficiency and the outcome of this human learning process is of high importance [184]. This research problem is widely considered in educational psychology through the lens of expectancy-value motivation theory. For instance, Hulleman et al. [174] propose to utilize *interventions* to increase the perception of usefulness (utility value) to subsequently increase motivation and final performance. Intervention here refers to identifying the relevance of model explanations to the user’s own situation, which can be a prompt question while working with the interface. Moreover, when utilizing model explanations in human-AI collaboration, explanations can be seen as a type of “scaffolding” (prompt during a task) proposed in a conceptual framework in education.

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

Hybrid Teaching. Teaching strategies for the human-to-human setting have been widely studied and many categorizations exist. One way of categorizing these strategies is through the following three concepts: (1) direct teaching, (2) indirect teaching, and (3) hybrid teaching. *Direct teaching* utilizes direct instructions that are teacher-centered, involve clear teaching objectives, and are consistent with classroom organizations. In XAI applications, direct teaching methods generate explanations by selecting representative examples of an agent’s decisions to convey the patterns in its policy. In contrast, *indirect teaching* is student-centered and encourages independent learning. In the XAI perspective, methods utilizing indirect teaching provide users with tools to actively and independently explore an AI system. Technically, direct teaching focuses on providing guidance (using a computational approach) to assist users in building an understanding of a machine, whereas indirect teaching (often through a user interface) enables users to address individual learning preferences and mitigate individual confusion about the AI. To leverage the advantages of the two teaching strategies, *hybrid teaching* has been widely used in human-to-human teaching with an emphasis on interactivity. Recent work [185] indicates that hybrid teaching reduces the amount of time for a user to understand an agent’s policy compared to direct and indirect teaching, and is more subjectively preferred by the participants. Building on this, future XAI systems can consider using hybrid teaching methods that (i) generate direct instructions to provide guidance to user’s understanding of an AI system; and (ii) provide methods to allow users to interact with the agent.

Explanations through Large Language Models (LLMs). The recent rise of Large Language Models [186, 187] naturally opens up new research directions. There is a growing interest in leveraging their unprecedented capabilities [188] to offer explanations for model decisions [189, 190]. Through their natural language interface, LLMs offer the possibility to build interactive explainers [191]. Intriguingly, textual explanations can also be used as subsequent inputs to LLMs which may help to solve subsequent problems and result in superior performance [192]. This technique, referred to as chain-of-thought reasoning [193], opens up an interesting research territory combining interpretability and performance considerations.

7.2 Open Research Problems

7.2.1 Automatic 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 [120, 121]. These metrics aim to test desiderata around the "faithfulness"/"fidelity"/ "truthfulness" of model explanations [121, 122, 194]. Faithfulness of explanations is defined as that explanations are indicative of true important features in the input [194]. The automatic evaluations aim at capturing general objectivity which is independent from downstream tasks, while human evaluations are contextualized with specific use cases. Generally speaking, automatic evaluations and human evaluations tackle different research challenges: the former objectively examines how truly explanations reflect models and the latter one measures how humans perceive models through explanations (although there existing algorithms for automated evaluation designed to align with human evaluations, which we will discuss later). All explanations used in human-subject experiments should have satisfying performance in automatic evaluations, i.e., the explanations should be able to faithfully unbox the model. This verification step is essential to guarantee the validity of the empirical user study and to ensure that users are not tricked by unfaithful explanations. However, in most current human-subject experiments, the functional faithfulness of explanations is not thoroughly verified beforehand. Using unfaithful explanations could lead to the problem that only the placebo effect of explanations is measured. Ideally, a good explanation should be faithful to the model as well as understandable by users.

7.2.2 Identifying and handling confounders

Existing research underscores the vulnerability of model explanation studies to significant confounding effects. For instance, Papenmeier et al. [158] reveal that user trust can be more influenced by model accuracy than the faithfulness of the explanation itself. Similarly, Yin et al. [195] demonstrate that the accuracy score perceived by users and the one shown to users contribute to trust formation.

A different problem is that good explanations also reveal weaknesses of the model. However, when seeing unexpected explanations, users may express their negative feelings about the model through negative ratings of the explanations. Therefore, good model explanations should help users *calibrate* their trust [26, 196], i.e., trust the model's decision when it is correct but distrust it otherwise. There is a disagreement on how to handle such cases: When evaluating model fairness, several works [2, 27, 38, 73, 75] reckon the increase in perceived fairness as positive, while Dodge et al. [74] define the decrease as positive. Other factors, such as the temporal occurrence of model errors (Nourani et al. [56]), and the dimensions of models (Ross et al. [32], Poursabzi et al. [53]), also come into play.

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

7.2.3 Mitigating personal biases for XAI

Most XAI techniques and corresponding designed user studies provide *one-size-fits-all* solutions. Individual bias, rooted in a user's mental framework, influences the user's perception of a model. It should be considered in XAI design, development, and evaluation procedures. Several studies that aim to explain reinforcement learning policies utilize cognitive science theories to create a model of the human user [184, 185, 197, 198]. They then generate explanations based on this human model and verify the benefits of tailoring explanations for individual user models. Within the scope of XAI, [199, 200] utilize a Bayesian Teaching framework to capture human perception of model explanations. In user studies, depending on cultural and educational background, participants may likely give different feedback [31]. This kind of personal bias can be mitigated by deploying a large sample size and recruiting participants who are representative of the target audience. We advocate that personal biases should be taken into account in the realm of XAI development.

7.2.4 Humans-in-the-Loop and sequential explanations

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

7.2.5 Proxy tasks should be close to real-world tasks

When using proxy tasks to evaluate models, for instance, to measure subjective understanding, there is a great choice of tasks present in the literature. A good proxy task should have the following features: (1) it has close real-world connections [120]; (2) users or participants have some background knowledge of the task but not too much to affect their judgment or performance during the task; (3) the task is not too complicated to implement or there exists an existing implementation but was used for different purposes (i.e., not used for XAI); and (4) it has connections to existing work. Yet, the link between evaluations through different proxy tasks and real-world applications has not been made very explicit to date. Buçinca et al. [13] show that the outcomes of proxy evaluations can be different from a real-world task. More specifically, the widely accepted proxy tasks, where users are asked to build the mental models of the AI, may not predict the performance in actual decision-making tasks, where users make use of the explanations to assist in making decisions. The results show that users trust different explanations in the proxy task and the actual decision-making task. Therefore, we argue that further research is required to uncover the links between current proxy tasks and on-task performance or to devise new proxy tasks with a verified connection to actual tasks.

7.2.6 Simulated evaluation as a cost-efficient solution

As human-subject experiments are costly to conduct, Chen et al. [201] propose a simulated evaluation framework (`SimEvals`) to select potential explanations for user studies by measuring the predictive information provided by explanations. Concretely, the authors consider three use cases where model explanations are deployed: forward simulation, counterfactual reasoning, and data debugging. Human performance is measured for these three tasks with different explanations. If there is a significant gap in settings of using two types of explanations, the simulated evaluation can also observe such a gap under the same task settings as well. Meanwhile, first attempts to simulate human textual responses in a given context using large language models show that models can provide surprisingly anthropomorphic answers [202]. Undoubtedly and also affirmed by Chen et al. [201], 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 [184, 185] that utilize this "Oz-of-Wizard" paradigm.

8 Conclusion

In recent years, there has been a proliferation of XAI research in both academia and industry. Explainability is a human-centric property [142] and therefore XAI should be preferably studied by taking humans' feedback into account. In this work, we investigated recent user studies for XAI techniques through a principled literature review. Based on our review, we found out that the effectiveness of XAI in users' interaction with ML models was not consistent across different applications, thus suggesting that there is a strong need for more transparent and comparable human-based evaluations in XAI. Furthermore, relevant disciplines, such as cognitive psychology and social sciences in general, should become an integral part of XAI research.

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

References

[1] C. Panigutti, A. Beretta, F. Giannotti, and D. Pedreschi, "Understanding the impact of explanations on advice-taking: a user study for ai-based clinical decision support systems," in *CHI*, 2022.
[2] A. I. Anik and A. Bunt, "Data-centric explanations: explaining training data of machine learning systems to promote transparency," in *CHI*, 2021.
[3] M. Colley, B. Eder, J. O. Rixen, and E. Rukzio, "Effects of semantic segmentation visualization on trust, situation awareness, and cognitive load in highly automated vehicles," in *CHI*, 2021.
[4] U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, and J. D. Weisz, "Expanding explainability: Towards social transparency in ai systems," in *CHI*, 2021.
[5] M. Liao and S. S. Sundar, "How should ai systems talk to users when collecting their personal information? effects of role framing and self-referencing on human-ai interaction," in *CHI*, 2021.
[6] C.-H. Tsai, Y. You, X. Gui, Y. Kou, and J. M. Carroll, "Exploring and promoting diagnostic transparency and explainability in online symptom checkers," in *CHI*, 2021.
[7] L. Guo, E. M. Daly, O. Alkan, M. Mattetti, O. Cornec, and B. Knijnenburg, "Building trust in interactive machine learning via user contributed interpretable rules," in *IUI*, 2022.
[8] J. Ooge, S. Kato, and K. Verbert, "Explaining recommendations in e-learning: Effects on adolescents' trust," in *IUI*, 2022.
[9] H. Suresh, K. M. Lewis, J. Guttag, and A. Satyanarayan, "Intuitively assessing ml model reliability through example-based explanations and editing model inputs," in *IUI*, 2022.
[10] R. Paleja, M. Ghuy, N. Ranawaka Arachchige, R. Jensen, and M. Gombolay, "The utility of explainable ai in ad hoc human-machine teaming," *NeurIPS*, 2021.
[11] J. Schaffer, J. O'Donovan, J. Michaelis, A. Raglin, and T. Höllerer, "I can do better than your ai: expertise and explanations," in *IUI*, 2019.
[12] X. Wang and M. Yin, "Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making," in *IUI*, 2021.
[13] Z. Buçinca, P. Lin, K. Z. Gajos, and E. L. Glassman, "Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems," in *IUI*, 2020.
[14] X. Peng, M. Riedl, and P. Ammanabrolu, "Inherently explainable reinforcement learning in natural language," in *NeurIPS*, 2022.
[15] Y. Zhang, Q. V. Liao, and R. K. Bellamy, "Effect of confidence and explanation on accuracy and trust cal-

ibration in ai-assisted decision making," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 295–305.

[16] V. Dominguez, P. Messina, I. Donoso-Guzmán, and D. Parra, "The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images," in *IUI*, 2019.

[17] C. J. Cai, J. Jongejan, and J. Holbrook, "The effects of example-based explanations in a machine learning interface," in *IUI*, 2019.

[18] M. Millecamp, N. N. Htun, C. Conati, and K. Verbert, "To explain or not to explain: the effects of personal characteristics when explaining music recommendations," in *IUI*, 2019.

[19] C.-H. Tsai and P. Brusilovsky, "Beyond the ranked list: User-driven exploration and diversification of social recommendation," in *IUI*, 2018.

[20] T. Li, G. Convertino, R. K. Tayi, and S. Kazerooni, "What data should i protect? recommender and planning support for data security analysts," in *IUI*, 2019.

[21] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan, "Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning," in *CHI*, 2020.

[22] H.-F. Cheng, R. Wang, Z. Zhang, F. O'Connell, T. Gray, F. M. Harper, and H. Zhu, "Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders," in *CHI*, 2019.

[23] J. Kunkel, T. Donkers, L. Michael, C.-M. Barbu, and J. Ziegler, "Let me explain: Impact of personal and impersonal explanations on trust in recommender systems," in *CHI*, 2019.

[24] D. H. Kim, E. Hoque, and M. Agrawala, "Answering questions about charts and generating visual explanations," in *CHI*, 2020.

[25] V. Lai and C. Tan, "On human predictions with explanations and predictions of machine learning models: A case study on deception detection," in *ACM FAccT*, 2019.

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

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

[28] U. Ehsan, P. Tambwekar, L. Chan, B. Harrison, and M. O. Riedl, "Automated rationale generation: a technique for explainable ai and its effects on human perceptions," in *IUI*, 2019.

[29] A. Smith-Renner, R. Fan, M. Birchfield, T. Wu, J. Boyd-Graber, D. S. Weld, and L. Findlater, "No explainability without accountability: An empirical study of explanations and feedback in interactive ml," in *CHI*, 2020.

[30] A. Smith-Renner, V. Kumar, J. Boyd-Graber, K. Seppi, and L. Findlater, "Digging into user control: perceptions of adherence and instability in transparent models," in *IUI*, 2020.

[31] A. Springer and S. Whittaker, "Progressive disclosure: empirically motivated approaches to designing effective transparency," in *IUI*, 2019.

[32] A. Ross, N. Chen, E. Z. Hang, E. L. Glassman, and F. Doshi-Velez, "Evaluating the interpretability of generative models by interactive reconstruction," in *CHI*, 2021.

[33] M. Radensky, D. Downey, K. Lo, Z. Popovic, and D. S. Weld, "Exploring the role of local and global explanations in recommender systems," in *CHI*, 2022.

[34] S. Hadash, M. C. Willemsen, C. Snijders, and W. A. IJsselsteijn, "Improving understandability of feature contributions in model-agnostic explainable ai tools," in *CHI*, 2022.

[35] M. Chromik, M. Eiband, F. Buchner, A. Krüger, and A. Butz, "I think i get your point, ai! the illusion of explanatory depth in explainable ai," in *IUI*, 2021.

[36] J. Rebanal, J. Combitsis, Y. Tang, and X. Chen, "Xalgo: A design probe of explaining algorithms' internal states via question-answering," in *IUI*, 2021.

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

[38] E. Rader, K. Cotter, and J. Cho, "Explanations as mechanisms for supporting algorithmic transparency," in *CHI*, 2018.

[39] A. Bell, I. Solano-Kamaiko, O. Nov, and J. Stoyanovich, "It's just not that simple: An empirical study of the accuracy-explainability trade-off in machine learning for public policy," in *ACM FAccT*, 2022.

[40] P. Hase and M. Bansal, "Evaluating explainable AI: Which algorithmic explanations help users predict model behavior?" in *ACL*, 2020.

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

[42] S. Bang, P. Xie, H. Lee, W. Wu, and E. Xing, "Explaining a black-box by using a deep variational information bottleneck approach," in *AAAI*, 2021.

[43] S. S. Kim, N. Meister, V. V. Ramaswamy, R. Fong, and O. Russakovsky, "Hive: Evaluating the human interpretability of visual explanations," in *ECCV*, 2022.

[44] M. Szymanski, M. Millecamp, and K. Verbert, "Visual, textual or hybrid: the effect of user expertise on different explanations," in *IUI*, 2021.

[45] G. Plumb, M. Al-Shedivat, Á. A. Cabrera, A. Perer, E. Xing, and A. Talwalkar, "Regularizing black-box models for improved interpretability," *NeurIPS*, 2020.

[46] W. Zhang and B. Y. Lim, "Towards relatable explainable ai with the perceptual process," in *CHI*, 2022.

[47] C. Bove, J. Aigrain, M.-J. Lesot, C. Tijus, and M. Detyniecki, "Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users," in *IUI*, 2022.

[48] A. Abdul, C. von der Weth, M. Kankanhalli, and B. Y. Lim, "Cogam: measuring and moderating cognitive load in machine learning model explanations," in *CHI*, 2020.

[49] K. Natesan Ramamurthy, B. Vinzamuri, Y. Zhang, and A. Dhurandhar, "Model agnostic multilevel explanations," *NeurIPS*, 2020.

[50] S. Arora, D. Pruthi, N. Sadeh, W. W. Cohen, Z. C. Lipton, and G. Neubig, "Explain, edit, and understand: Rethinking user study design for evaluating model explanations," in *AAAI*, 2022.

[51] J. Antoran, U. Bhatt, T. Adel, A. Weller, and J. M. Hernández-Lobato, "Getting a {clue}: A method for explaining uncertainty estimates," in *ICLR*, 2021.

[52] J. Borowski, R. S. Zimmermann, J. Schepers, R. Geirhos, T. S. A. Wallis, M. Bethge, and W. Brendel, "Exemplary natural images explain {cnn} activations better than state-of-the-art feature visualization," in *ICLR*, 2021.

[53] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, and H. Wallach, "Manipulating and measuring model interpretability," in *CHI*, 2021.

[54] A. Alqaraawi, M. Schuessler, P. Weiß, E. Costanza, and N. Berthouze, "Evaluating saliency map explanations for convolutional neural networks: a user study," in *IUI*, 2020.

[55] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *AAAI*, 2018.

[56] M. Nourani, C. Roy, J. E. Block, D. R. Honeycutt, T. Rahman, E. Ragan, and V. Gogate, "Anchoring bias affects mental model formation and user reliance in explainable ai systems," in *IUI*, 2021.

[57] L. Sixt, M. Schuessler, O.-I. Popescu, P. Weiß, and T. Landgraf, "Do users benefit from interpretable vision? a user study, baseline, and dataset," in *ICLR*, 2022.

[58] A. Chandrasekaran, V. Prabhu, D. Yadav, P. Chattopadhyay, and D. Parikh, "Do explanations make vqa models more predictable to a human?" in *EMNLP*, 2018.

[59] J. Colin, T. Fel, R. Cadene, and T. Serre, "What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods," in *NeurIPS*, 2022.

[60] H. Shen and T.-H. Huang, "How useful are the machine-generated interpretations to general users? a human evaluation on guessing the incorrectly predicted labels," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 8, no. 1, 2020, pp. 168–172.

[61] C.-K. Yeh, B. Kim, S. O. Arik, C.-L. Li, T. Pfister, and P. Ravikumar, "On completeness-aware concept-based explanations in deep neural networks," in *NeurIPS*, 2019.

[62] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, "Towards automatic concept-based explanations," in *NeurIPS*, 2019.

[63] T. Leemann, Y. Rong, S. Kraft, E. Kasneci, and G. Kasneci, "Coherence evaluation of visual concepts with objects and language," in *ICLR2022 WS*, 2022.

[64] I. Laina, R. Fong, and A. Vedaldi, "Quantifying learnability and describability of visual concepts emerging in representation learning," 2020.

[65] Y. Wang, P. Venkatesh, and B. Y. Lim, "Interpretable directed diversity: Leveraging model explanations for iterative crowd ideation," in *CHI*, 2022.

[66] D. L. Arendt, N. Nur, Z. Huang, G. Fair, and W. Dou, "Parallel embeddings: a visualization technique for contrasting learned representations," in *IUI*, 2020.

[67] W. Zhang, M. Dimiccoli, and B. Y. Lim, "Debiased-cam to mitigate image perturbations with faithful visual explanations of machine learning," in *CHI*, 2022.

[68] J. Gao, X. Wang, Y. Wang, and X. Xie, "Explainable recommendation through attentive multi-view learning," in *AAAI*, 2019.

[69] P. Kouki, J. Schaffer, J. Pujara, J. O'Donovan, and L. Getoor, "Personalized explanations for hybrid recommender systems," in *IUI*, 2019.

[70] C.-H. Tsai and P. Brusilovsky, "Explaining recommendations in an interactive hybrid social recommender," in *IUI*, 2019.

[71] A. Balayn, N. Rikalo, C. Lofi, J. Yang, and A. Bozzon, "How can explainability methods be used to support bug identification in computer vision models?" in *CHI*, 2022.

[72] K. Rawal and H. Lakkaraju, "Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses," *NeurIPS*, 2020.

[73] N. Grgić-Hlača, E. M. Redmiles, K. P. Gummadi, and A. Weller, "Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction," in *WWW*, 2018.

[74] J. Dodge, Q. V. Liao, Y. Zhang, R. K. Bellamy, and C. Dugan, "Explaining models: an empirical study of how explanations impact fairness judgment," in *IUI*, 2019.

[75] G. Harrison, J. Hanson, C. Jacinto, J. Ramirez, and B. Ur, "An empirical study on the perceived fairness of realistic, imperfect machine learning models," in *ACM FAccT*, 2020.

[76] C. Wang, K. Wang, A. Bian, R. Islam, K. N. Keya, J. Foulds, and S. Pan, "Do humans prefer debiased ai algorithms? a case study in career recommendation," in *IUI*, 2022.

[77] N. N. Htun, E. Lecluse, and K. Verbert, "Perception of fairness in group music recommender systems," in *IUI*, 2021.

[78] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt, "'it's reducing a human being to a percentage' perceptions of justice in algorithmic decisions," in *CHI*, 2018.

[79] J. Schoeffer and N. Kuehl, "Appropriate fairness perceptions? on the effectiveness of explanations in enabling people to assess the fairness of automated decision systems," in *CSCW*, 2021.

[80] T. Donkers, T. Kleemann, and J. Ziegler, "Explaining recommendations by means of aspect-based transparent memories," in *IUI*, 2020.

[81] F. Hohman, A. Head, R. Caruana, R. DeLine, and S. M. Drucker, "Gamut: A design probe to understand how data scientists understand machine learning models," in *CHI*, 2019.

[82] U. Kuhl, A. Artelt, and B. Hammer, "Let's go to the alien zoo: Introducing an experimental framework to

study usability of counterfactual explanations for machine learning," *arXiv preprint arXiv:2205.03398*, 2022.

[83] T. Schneider, J. Hois, A. Rosenstein, S. Ghellal, D. Theofanou-Fülbier, and A. R. Gerlicher, "Explain yourself! transparency for positive ux in autonomous driving," in *CHI*, 2021.

[84] S. Choi, K. Aizawa, and N. Sebe, "Fontmatcher: font image paring for harmonious digital graphic design," in *IUI*, 2018.

[85] P. Le Bras, D. A. Robb, T. S. Methven, S. Padilla, and M. J. Chantler, "Improving user confidence in concept maps: Exploring data driven explanations," in *CHI*, 2018.

[86] R. Shang, K. K. Feng, and C. Shah, "Why am i not seeing it? understanding users' needs for counterfactual explanations in everyday recommendations," in *ACM FAccT*, 2022.

[87] J. Dodge, A. A. Anderson, M. Olson, R. Dikkala, and M. Burnett, "How do people rank multiple mutant agents?" in *IUI*, 2022.

[88] D. Das and S. Chernova, "Leveraging rationales to improve human task performance," in *IUI*, 2020.

[89] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld, "Does the whole exceed its parts? the effect of ai explanations on complementary team performance," in *CHI*, 2021.

[90] V. Lai, H. Liu, and C. Tan, "" why is' chicago'deceptive?" towards building model-driven tutorials for humans," in *CHI*, 2020.

[91] S. Feng and J. Boyd-Graber, "What can ai do for me? evaluating machine learning interpretations in cooperative play," in *IUI*, 2019.

[92] Y. Alufaisan, L. R. Marusich, J. Z. Bakdash, Y. Zhou, and M. Kantarcioglu, "Does explainable artificial intelligence improve human decision-making?" in *AAAI*, 2021.

[93] K. Z. Gajos and L. Mamykina, "Do people engage cognitively with ai? impact of ai assistance on incidental learning," in *IUI*, 2022.

[94] M. Liao, S. S. Sundar, and J. B. Walther, "User trust in recommendation systems: A comparison of content-based, collaborative and demographic filtering," in *CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–14.

[95] G. Nguyen, D. Kim, and A. Nguyen, "The effectiveness of feature attribution methods and its correlation with automatic evaluation scores," in *NeurIPS*, 2021.

[96] M. R. Taesiri, G. Nguyen, and A. Nguyen, "Visual correspondence-based explanations improve AI robustness and human-AI team accuracy," in *NeurIPS*, 2022.

[97] G. Nguyen, M. R. Taesiri, and A. Nguyen, "Visual correspondence-based explanations improve ai robustness and human-ai team accuracy," *NeurIPS*, 2022.

[98] J. Wei, J. He, K. Chen, Y. Zhou, and Z. Tang, "Collaborative filtering and deep learning based recommendation system for cold start items," *Expert Systems with Applications*, 2017.

[99] S. Yang, M. Korayem, K. AlJadda, T. Grainger, and S. Natarajan, "Combining content-based and collaborative filtering for job recommendation system: A cost-sensitive statistical relational learning approach," *Knowledge-Based Systems*, 2017.

[100] Y. Zhang, X. Chen, Q. Ai, L. Yang, and W. B. Croft, "Towards conversational search and recommendation: System ask, user respond," in *CIKM*, 2018.

[101] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *Journal of Field Robotics*, 2020.

[102] H. Cui, V. Radosavljevic, F.-C. Chou, T.-H. Lin, T. Nguyen, T.-K. Huang, J. Schneider, and N. Djuric, "Multimodal trajectory predictions for autonomous driving using deep convolutional networks," in *ICR*, 2019.

[103] Y. Rong, C. Han, C. Hellert, A. Loyal, and E. Kasneci, "Artificial intelligence methods in in-cabin use cases: a survey," *IEEE ITSM*, 2021.

[104] R. R. Murphy, *Introduction to AI robotics*, 2019.

[105] K. Rajan and A. Saffiotti, "Towards a science of integrated ai and robotics," 2017.

[106] S. Wachter, B. Mittelstadt, and L. Floridi, "Transparent, explainable, and accountable ai for robotics," *Science robotics*, 2017.

[107] S. H. Park and K. Han, "Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction," *Radiology*, 2018.

[108] J. A. Sidey-Gibbons and C. J. Sidey-Gibbons, "Machine learning in medicine: a practical introduction," *BMC medical research methodology*, 2019.

[109] R. Vaishya, M. Javaid, I. H. Khan, and A. Haleem, "Artificial intelligence (ai) applications for covid-19 pandemic," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 2020.

[110] X. Dastile, T. Celik, and M. Potsane, "Statistical and machine learning models in credit scoring: A systematic literature survey," *Applied Soft Computing*, 2020.

[111] M. Ala'raj, M. F. Abbod, M. Majdalawieh, and L. Jum'a, "A deep learning model for behavioural credit scoring in banks," *Neural Computing and Applications*, 2022.

[112] P. M. Addo, D. Guegan, and B. Hassani, "Credit risk analysis using machine and deep learning models," *Risks*, 2018.

[113] N. Van Berkel, J. Goncalves, D. Hettiachchi, S. Wijenayake, R. M. Kelly, and V. Kostakos, "Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study," *PACM HCI*, 2019.

[114] T. Sourdin, "Judge v robot?: Artificial intelligence and judicial decision-making," *University of New South Wales Law Journal, The*, 2018.

[115] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic hiring: Evaluating claims and practices," in *ACM FAccT*, 2020.

[116] P. Tambe, P. Cappelli, and V. Yakubovich, "Artificial intelligence in human resources management: Challenges and a path forward," *California Management Review*, 2019.

[117] D. Castelvecchi, "Can we open the black box of ai?" *Nature News*, 2016.

[118] M. O. Riedl, "Human-centered artificial intelligence

and machine learning," *Human Behavior and Emerging Technologies*, 2019.
[119] U. Ehsan and M. O. Riedl, "Human-centered explainable ai: Towards a reflective sociotechnical approach," in *HCII*, 2020.
[120] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.
[121] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, and C. Seifert, "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai," *ACM Computing Surveys*, 2023.
[122] R. Tomsett, D. Harborne, S. Chakraborty, P. Gurram, and A. Preece, "Sanity checks for saliency metrics," in *AAAI*, 2020.
[123] Y. Rong, T. Leemann, V. Borisov, G. Kasneci, and E. Kasneci, "A consistent and efficient evaluation strategy for attribution methods," in *ICML*, 2022.
[124] D. Nguyen, "Comparing automatic and human evaluation of local explanations for text classification," in *NAACL HLT*, 2018.
[125] G. Hoffman, "Evaluating fluency in human–robot collaboration," *IEEE Transactions on Human-Machine Systems*, 2019.
[126] M. Chromik and M. Schuessler, "A taxonomy for human subject evaluation of black-box explanations in xai." *Exss-atec@ iui*, 2020.
[127] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable ai systems," *TiiS*, 2021.
[128] Q. Yang, N. Banovic, and J. Zimmerman, "Mapping machine learning advances from hci research to reveal starting places for design innovation," in *CHI*, 2018.
[129] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," *IEEE access*, 2018.
[130] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, 2020.
[131] W. Samek and K.-R. Müller, "Towards explainable artificial intelligence," in *Explainable AI: interpreting, explaining and visualizing deep learning*, 2019.
[132] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *Journal of Artificial Intelligence Research*, 2021.
[133] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, 2019.
[134] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *DSAA*, 2018.
[135] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, "Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda," in *CHI*, 2018.
[136] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital signal processing*, 2018.
[137] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (xai): A survey," *arXiv preprint arXiv:2006.11371*, 2020.
[138] G. Joshi, R. Walambe, and K. Kotecha, "A review on explainability in multimodal deep neural nets," *IEEE Access*, 2021.
[139] R. Moraffah, M. Karami, R. Guo, A. Raglin, and H. Liu, "Causal interpretability for machine learning-problems, methods and evaluation," *ACM SIGKDD Explorations Newsletter*, 2020.
[140] I. Nunes and D. Jannach, "A systematic review and taxonomy of explanations in decision support and recommender systems," *UMUAI*, 2017.
[141] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, 2018.
[142] Q. V. Liao and K. R. Varshney, "Human-centered explainable ai (xai): From algorithms to user experiences," *arXiv preprint arXiv:2110.10790*, 2021.
[143] V. Lai, C. Chen, Q. V. Liao, A. Smith-Renner, and C. Tan, "Towards a science of human-ai decision making: a survey of empirical studies," *arXiv preprint arXiv:2112.11471*, 2021.
[144] J. J. Ferreira and M. S. Monteiro, "What are people doing about xai user experience? a survey on ai explainability research and practice," in *HCII*, 2020.
[145] I. S. . E. of Human-System Interaction (Subcommittee), *Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs).: Guidance on Usability*, 1998.
[146] S. A. Cholewiak, P. Ipeirotis, V. Silva, and A. Kannawadi, "SCHOLARLY: Simple access to Google Scholar authors and citation using Python," 2021.
[147] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *NeurIPS*, 2020.
[148] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, 2008.
[149] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *KDD*, 2016.
[150] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *NeurIPS*, 2017.
[151] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017.
[152] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.
[153] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI magazine*, 2017.
[154] C. Molnar, *Interpretable Machine Learning*, 2019.
[155] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, 2019.
[156] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm,

and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *KDD*, 2015.
[157] C. Panigutti, A. Perotti, and D. Pedreschi, "Doctor xai: an ontology-based approach to black-box sequential data classification explanations," in *ACM FAccT*, 2020.
[158] A. Papenmeier, G. Englebienne, and C. Seifert, "How model accuracy and explanation fidelity influence user trust," *arXiv preprint arXiv:1907.12652*, 2019.
[159] J. van der Waa, E. Nieuwburg, A. Cremers, and M. Neerincx, "Evaluating xai: A comparison of rule-based and example-based explanations," *Artificial Intelligence*, 2021.
[160] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine learning for medical imaging," *Radiographics*, 2017.
[161] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *International journal of cognitive ergonomics*, 2000.
[162] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *ICML*, 2018.
[163] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell, "Explaining the user experience of recommender systems," *User modeling and user-adapted interaction*, 2012.
[164] B. Y. Lim and A. K. Dey, "Assessing demand for intelligibility in context-aware applications," in *UbiComp*, 2009.
[165] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in psychology*, 1988.
[166] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: Challenges and prospects," *arXiv preprint arXiv:1812.04608*, 2018.
[167] A. Holzinger, A. Carrington, and H. Müller, "Measuring the quality of explanations: the system causability scale (scs)," *KI-Künstliche Intelligenz*, 2020.
[168] A. Gegenfurtner, E. Lehtinen, and R. Säljö, "Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains," *Educational psychology review*, 2011.
[169] K. Cotter, J. Cho, and E. Rader, "Explaining the news feed algorithm: An analysis of the" news feed fyi" blog," in *Extended abstracts at CHI*, 2017.
[170] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, "Designing theory-driven user-centric explainable ai," in *CHI*, 2019.
[171] L. Rozenblit and F. Keil, "The misunderstood limits of folk science: An illusion of explanatory depth," *Cognitive science*, 2002.
[172] G. Hoffman and X. Zhao, "A primer for conducting experiments in human–robot interaction," *THRI*, 2020.
[173] J. Eccles, "Expectancies, values and academic behaviors," *Achievement and achievement motives*, 1983.
[174] C. S. Hulleman, J. J. Kosovich, K. E. Barron, and D. B. Daniel, "Making connections: Replicating and extending the utility value intervention in the classroom." *Journal of Educational Psychology*, 2017.
[175] F. G. Paas, "Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach." *Journal of educational psychology*, 1992.
[176] K. Ouwehand, A. v. d. Kroef, J. Wong, and F. Paas, "Measuring cognitive load: Are there more valid alternatives to likert rating scales?" in *Frontiers in Education*, 2021.
[177] J. P. Simmons, L. D. Nelson, and U. Simonsohn, "Pre-registration: Why and how," *Journal of Consumer Psychology*, 2021.
[178] U. Simonsohn, L. D. Nelson, and J. P. Simmons, "P-curve: a key to the file-drawer." *Journal of experimental psychology: General*, 2014.
[179] K. A. Ericsson and H. A. Simon, *Protocol analysis: Verbal reports as data.*, 1984.
[180] J. Cohen, *Statistical power analysis for the behavioral sciences*, 2013.
[181] S. Dhanorkar, C. T. Wolf, K. Qian, A. Xu, L. Popa, and Y. Li, "Who needs to know what, when?: Broadening the explainable ai (xai) design space by looking at explanations across the ai lifecycle," in *Designing Interactive Systems Conference 2021*, 2021, pp. 1591–1602.
[182] F. Y. Kung, N. Kwok, and D. J. Brown, "Are attention check questions a threat to scale validity?" *Applied Psychology*, 2018.
[183] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, 1971.
[184] I. Lage, D. Lifschitz, F. Doshi-Velez, and O. Amir, "Exploring computational user models for agent policy summarization," in *IJCAI*, 2019.
[185] P. Qian and V. Unhelkar, "Evaluating the role of interactivity on improving transparency in autonomous agents," in *AAMAS*, 2022.
[186] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
[187] OpenAI, "optimizing language models for dialogue. technical report." 2023.
[188] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg *et al.*, "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, 2023.
[189] W. Zhou, J. Hu, H. Zhang, X. Liang, M. Sun, C. Xiong, and J. Tang, "Towards interpretable natural language understanding with explanations as latent variables," *NeurIPS*, 2020.
[190] S. Wiegreffe, J. Hessel, S. Swayamdipta, M. Riedl, and Y. Choi, "Reframing human-ai collaboration for generating free-text explanations," in *NAACL-HLT*, 2022.
[191] S. Wang, Z. Zhao, X. Ouyang, Q. Wang, and D. Shen, "Chatcad: Interactive computer-aided diagnosis on medical image using large language models," *arXiv preprint arXiv:2302.07257*, 2023.
[192] N. F. Rajani, B. McCann, C. Xiong, and R. Socher, "Explain yourself! leveraging language models for commonsense reasoning," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 4932–4942.
[193] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia,

E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *NeurIPS*, 2022.
[194] D. Alvarez Melis and T. Jaakkola, "Towards robust interpretability with self-explaining neural networks," *NeurIPS*, 2018.
[195] M. Yin, J. Wortman Vaughan, and H. Wallach, "Understanding the effect of accuracy on trust in machine learning models," in *CHI*, 2019.
[196] A. Bussone, S. Stumpf, and D. O'Sullivan, "The role of explanations on trust and reliance in clinical decision support systems," in *ICHI*, 2015.
[197] C. Baker, R. Saxe, and J. Tenenbaum, "Bayesian theory of mind: Modeling joint belief-desire attribution," in *Proceedings of the annual meeting of the cognitive science society*, 2011.
[198] S. H. Huang, D. Held, P. Abbeel, and A. D. Dragan, "Enabling robots to communicate their objectives," *Autonomous Robots*, 2019.
[199] S. C.-H. Yang, N. E. T. Folke, and P. Shafto, "A psychological theory of explainability," in *ICML*, 2022.
[200] S. C.-H. Yang, W. K. Vong, R. B. Sojitra, T. Folke, and P. Shafto, "Mitigating belief projection in explainable artificial intelligence via bayesian teaching," *Scientific reports*, 2021.
[201] V. Chen, N. Johnson, N. Topin, G. Plumb, and A. Talwalkar, "Use-case-grounded simulations for explanation evaluation," *arXiv preprint arXiv:2206.02256*, 2022.
[202] G. Aher, R. I. Arriaga, and A. T. Kalai, "Using large language models to simulate multiple humans," *arXiv preprint arXiv:2208.10264*, 2022.
[203] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, 2019.
[204] Q. V. Liao, M. Pribić, J. Han, S. Miller, and D. Sow, "Question-driven design process for explainable ai user experiences," *arXiv preprint arXiv:2104.03483*, 2021.
[205] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, "Principles of explanatory debugging to personalize interactive machine learning," in *IUI*, 2015.
[206] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W.-K. Wong, "Too much, too little, or just right? ways explanations impact end users' mental models," in *VL/HCC*, 2013.
[207] T. Kulesza, S. Stumpf, M. Burnett, and I. Kwan, "Tell me more? the effects of mental model soundness on personalizing an intelligent agent," in *CHI*, 2012.
[208] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *CSUR*, 2018.
[209] J. L. Herlocker, J. A. Konstan, and J. Riedl, "Explaining collaborative filtering recommendations," in *CSCW*, 2000.
[210] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *ICML*, 2017.
[211] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, 2017.
[212] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *ICML*, 2017.
[213] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.
[214] C. J. Cai, E. Reif, N. Hegde, J. Hipp, B. Kim, D. Smilkov, M. Wattenberg, F. Viegas, G. S. Corrado, M. C. Stumpe *et al.*, "Human-centered tools for coping with imperfect algorithms during medical decision-making," in *CHI*, 2019.
[215] J. Krause, A. Perer, and K. Ng, "Interacting with predictions: Visual inspection of black-box machine learning models," in *CHI*, 2016.
[216] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," *NeurIPS*, 2016.
[217] B. Y. Lim, A. K. Dey, and D. Avrahami, "Why and why not explanations improve the intelligibility of context-aware intelligent systems," in *CHI*, 2009.
[218] M. Narayanan, E. Chen, J. He, B. Kim, S. Gershman, and F. Doshi-Velez, "How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation," *arXiv preprint arXiv:1802.00682*, 2018.
[219] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human factors*, 2004.
[220] R. F. Kizilcec, "How much information? effects of transparency on trust in an algorithmic interface," in *CHI*, 2016.
[221] H. Cramer, V. Evers, S. Ramlal, M. Van Someren, L. Rutledge, N. Stash, L. Aroyo, and B. Wielinga, "The effects of transparency on trust in and acceptance of a content-based art recommender," *User Modeling and User-adapted interaction*, 2008.
[222] G. Friedrich and M. Zanker, "A taxonomy for generating explanations in recommender systems," *AI Magazine*, 2011.
[223] P. Kouki, J. Schaffer, J. Pujara, J. O'Donovan, and L. Getoor, "User preferences for hybrid explanations," in *RecSys*, 2017.
[224] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, 2015.
[225] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.
[226] J.-H. Jacobsen, A. W. Smeulders, and E. Oyallon, "i-revnet: Deep invertible networks," in *ICLR*, 2018.
[227] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: deep learning for interpretable image recognition," *NeurIPS*, 2019.
[228] K. Natesan Ramamurthy, B. Vinzamuri, Y. Zhang, and A. Dhurandhar, "Model agnostic multilevel explanations," *Advances in neural information processing systems*, vol. 33, pp. 5968–5979, 2020.
[229] A. Balayn, P. Soilis, C. Lofi, J. Yang, and A. Bozzon, "What do you mean? interpreting image classification with crowdsourced concept extraction and analysis," in *WWW*, 2021.
[230] M. Madsen and S. Gregor, "Measuring human-

computer trust," in *11th australasian conference on information systems*, 2000.

[231] I. Benbasat and W. Wang, "Trust in and adoption of online recommendation agents," *Journal of the association for information systems*, 2005.

[232] J. Lee and N. Moray, "Trust, control strategies and allocation of function in human-machine systems," *Ergonomics*, 1992.

[233] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS quarterly*, 2003.

[234] D. H. McKnight, V. Choudhury, and C. Kacmar, "Developing and validating trust measures for e-commerce: An integrative typology," *Information systems research*, 2002.

[235] C. Hewitt, I. Politis, T. Amanatidis, and A. Sarkar, "Assessing public perception of self-driving cars: the autonomous vehicle acceptance model," in *IUI*, 2019.

[236] M. Schrepp, A. Hinderks, and J. Thomaschewski, "Design and evaluation of a short version of the user experience questionnaire (ueq-s)," *IJIMAI*, 2017.

[237] J. Sauro and J. S. Dumas, "Comparison of three one-question, post-task usability questionnaires," in *CHI*, 2009.

[238] H. L. O'Brien, P. Cairns, and M. Hall, "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*, 2018.

[239] J. Brooke *et al.*, "Sus-a quick and dirty usability scale," *Usability evaluation in industry*, 1996.

[240] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," in *ICML*, 2020.

[241] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020.

[242] A. YM., R. C., and V. A., "Self-labelling via simultaneous clustering and representation learning," in *ICLR*, 2020.

[243] M. Richardson, C. Abraham, and R. Bond, "Psychological correlates of university students' academic performance: a systematic review and meta-analysis." *Psychological bulletin*, 2012.

[244] A. Wigfield and J. Cambria, "Expectancy-value theory: Retrospective and prospective," in *The decade ahead: Theoretical perspectives on motivation and achievement*, 2010.

[245] O. Chernikova, N. Heitzmann, A. Opitz, T. Seidel, and F. Fischer, "A theoretical framework for fostering diagnostic competences with simulations in higher education," *Learning to Diagnose with Simulations*, 2022.

[246] N. Heitzman, T. Seidel, A. Opitz, A. Hetmanek, C. Wecker, M. Fischer, S. Ufer, R. Schmidmaier, B. Neuhaus, M. Siebeck *et al.*, "Facilitating diagnostic competences in simulations: A conceptual framework and a research agenda for medical and teacher education." *Frontline Learning Research*, 2019.

[247] K. Bisra, Q. Liu, J. C. Nesbit, F. Salimi, and P. H. Winne, "Inducing self-explanation: A meta-analysis," *Educational Psychology Review*, 2018.

[248] C. Conati, O. Barral, V. Putnam, and L. Rieger, "Toward personalized xai: A case study in intelligent tutoring systems," *Artificial Intelligence*, 2021.

[249] G. S. Becker, *The economic approach to human behavior*, 1976.

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

[251] C. L. Baker, "Bayesian theory of mind: Modeling human reasoning about beliefs, desires, goals, and social relations," Ph.D. dissertation, 2012.

[252] G. Csibra, "Cognitive science: Modelling theory of mind," *Nature Human Behaviour*, 2017.

[253] T. Hellström and S. Bensch, "Understandable robots," *Paladyn, Journal of Behavioral Robotics*, 2018.

[254] S. lai Lee, I. Y. man Lau, S. Kiesler, and C.-Y. Chiu, "Human mental models of humanoid robots," in *ICRA*, 2005.

[255] M. S. Lee, H. Admoni, and R. Simmons, "Machine teaching for human inverse reinforcement learning," *Frontiers in Robotics and AI*, 2021.

[256] L. Julien-Schultz, N. Maynes, and C. Dunn, "Managing direct and indirect instruction: A visual model to support lesson planning in pre-service programs," *The International Journal of Learning: Annual Review*, 2010.

[257] K. A. Nguyen, J. Husman, M. A. T. Borrego, P. Shekhar, M. J. Prince, M. DeMonbrun, C. J. Finelli, C. Henderson, and C. K. Waters, "Students' expectations, types of instruction, and instructor strategies predicting student response to active learning," *IJEE*, 2017.

[258] T. Ruutmann and H. Kipper, "Teaching strategies for direct and indirect instruction in teaching engineering," in *2011 14th International Conference on Interactive Collaborative Learning*, 2011.

[259] D. Amir and O. Amir, "Highlights: Summarizing agent behavior to people," in *AAMAS*, 2018.

[260] S. H. Huang, K. Bhatia, P. Abbeel, and A. D. Dragan, "Establishing appropriate trust via critical states," *IROS*, 2018.

[261] O. Amir, F. Doshi-Velez, and D. Sarne, "Summarizing agent strategies," *Autonomous Agents and Multi-Agent Systems*, 2019.

[262] O. Watkins, S. Huang, J. Frost, K. Bhatia, E. Weiner, P. Abbeel, T. Darrell, B. Plummer, K. Saenko, and A. Dragan, "Explaining robot policies," *Applied AI Letters*, 2021.

[263] O. A. Yotam Amitai, ""i don't think so": Summarizing policy disagreements for agent comparison," 2022.

[264] C. P. Fulford and S. Zhang, "Perceptions of interaction: The critical predictor in distance education," *American Journal of Distance Education*, 1993.

[265] "Interactivity research studies," *Journal of Educational Technology & Society*, 2001.

[266] B. Muirhead and C. Juwah, "Interactivity in computer-mediated college and university education: A recent review of the literature," *Educational Technology & Society*, 2004.

APPENDIX 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 6, we categorize these papers according to their topics. The first group of papers is survey papers about XAI, which are thoroughly discussed in Sec.2 Related Work. For the theory of XAI, Miller et al. [203] propose to build XAI on social sciences such as cognitive science and psychology, while Wang et al. [170] and Liao et al. [204] 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 [205, 206], the explanations should be sound and complete and thus bring positive impact on users. Another motivation for XAI is that it should assist users in building mental models of the AI systems [207]. 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.

APPENDIX B
MODELS AND EXPLANATIONS IN XAI USER STUDIES

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 [149] and SHAP [150]. Figure 4 demonstrates the chronological overview of frequently adopted XAI techniques for black-box models in user studies from the surveyed papers. However, there are many specific explanation types for certain applications. 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 [69, 222, 223]. For instance, Dominguez et al. [16] provide a content-based explanation as *"Painting A is 85% similar to the Painting B that you like"*. Tsai et al. [70], however, use hybrid explanations in textual and visual explanation formats.

APPENDIX C
MEASUREMENT DETAILS

C.1 Trust

Table 8 lists the trust measurement. Most of the works deploy questionnaires to measure user trust (self-reported), where a 7-point or 5-point Likert scale is commonly used. Many works design their own questionnaires [6, 11, 13, 16, 17, 18, 19, 24, 27]. To measure trust in an objective manner, many works choose to use the agreement rate of humans [9, 11, 12, 25].

C.2 Usability

Table 9 demonstrates the measures used for the usability of explanations. We divide usability into five sub-categories: workload (cognitive load), helpfulness, satisfaction, undesired behavior detection 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 the accuracy of the user confirming the correctness of answers from a question-answering model and the time for solving this task [24].

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 [162, 240], such kind of understanding questions are common, to assess semantic coherence of automatically discovered concepts [61, 62, 63, 64]. Assignment tasks, where novel instances should be assigned to existing clusters are commonly used as a proxy to measure the intelligibility [49, 61, 62, 64]. Another option is to assess how well the cluster can be described in natural language which is often referred to as *describability* [62, 63, 64]. Apart from conceptual explanations, Zhang et al. [46] ask multiple choice questions to verify if users understand the differences between the acoustical cues presented and evaluate which cue differences were most noticeable. Wang et al. [65] prompt users explicitly if the found the explanation easy to understand.

Research questions and Findings. Laina et al. [64] found that feature vectors obtained by contrastive learning approaches such as MoCo [241] or SeLa [242] allow for clusters that are almost as interpretable as human labels. Leemann et al. [63] show the similarity of ResNet-50 embeddings allows to predict how semantically coherent users find a cluster of images. For the acoustical cue, Zhang et al. [46] found that shrillness and speaking rate were most often recognized. Wang et al. [12] found that users reported they understood all types of explanations well without significant differences.

APPENDIX D
FINDINGS

When using explanation types as the evaluation dimension, many works compare their effects without comparing them to a control group (baseline) without explanation methods. Anik et al. [2] argue that many works have proven the usefulness of explanations and therefore no need to include such a control group. Table 11 summarizes the findings of the comparison among different explanations. Table 10 lists results of using other evaluation dimensions beyond explanations.

Topic	Fundamental works
Surveys of XAI	[120], [141], [135], [129], [134], [125], [130]
Theories for XAI	[203]: social sciences, [170]: theory for XAI design, [204]: a question bank for XAI design
XAI Methods	[208]: a survey, [149]: LIME, [55]: Anchors, [150]: SHAP, [162]: TCAV, [209]: explaining recommendation systems, [156]: intelligible models, [210]: influence function, [211]: counterfactual explanations, [212]: Integrated Gradient (IG), [213]: saliency maps for images, [151]: GradCAM
Principles of Explanations	[205, 206]: completeness and soundness, [207]: helping users build mental models
User studies for ML	[214]: image retrieval algorithm for medical uses, [215]: interactive model
User studies for XAI	[78]: justice perceptions, [74]: fairness [25]: human-AI team, [81]: usability, [38, 53, 216, 217, 218]: understanding, [13, 17, 21, 22]: trust and understanding
Trust	[196]: trust (calibration), [219]: trust in automation, [195]: impact of model accuracy on trust, [220, 221]: impact of system transparency on trust,

TABLE 6: Fundamental works of the core papers (categorized according to topics).

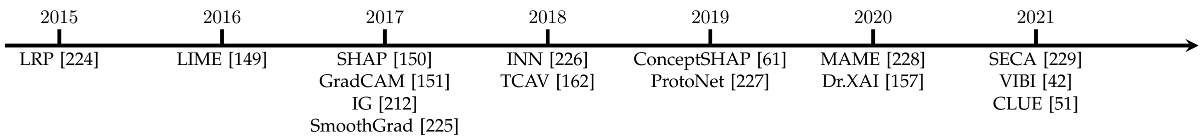

Fig. 4: Chronology of commonly used XAI methods from reviewed papers.

Tasks	Tabular	Image/Video	Text	Other
forward simulation	[40, 49, 53] [35, 51, 55] [12, 22, 39]	[13, 52, 56] [54, 58, 59] [43, 60][55, (VQA)]	[40, 50]	[46] (Audio)
marginal feature effects	[40, 47, 48] [12, 22]		[40]	
manipulation / counterfactual sim.	[12, 22, 45]	[32]	[50]	
feature importance	[12, 39, 47]	[57, 162]		
failure prediction		[58]		
relative simulation (selection)	[22, 35]			
other	[48] (mental model faithfulness)	[56] (class-wise acc.)		

TABLE 7: Works measuring objective understanding grouped by proxy task/data modality

APPENDIX E
TOWARDS INCREASINGLY USER-CENTERED XAI

In this section, we provide a detailed literature review regarding existing work in pedagogical frameworks, which provides implications for designing future transparent AI systems and human-centered evaluations in Sec. 7.1.

E.1 Expectancy-value Motivation Theory

Human interaction with XAI interfaces can be viewed as an activity where humans learn about the model's inner workings through explanations and then achieve an understanding of the models. The question of how to enhance the efficiency and the outcome of this human learning process is of high importance [184]. This research question is widely considered in educational psychology through the lens of expectancy-value motivation theory [174, 243, 244]. For instance, Hulleman et al. [174] propose to utilize *interventions* to increase the perception of usefulness (utility value) to subsequently increase motivation and final performance. Intervention here refers to identifying the relevance of model explanations to the user's own situation, which can be a prompt question while working with the interface. Moreover, when utilizing model explanations in human-AI collaboration, explanations can be seen as a type of "scaffolding" (prompt during a task) proposed in a conceptual framework in education [245, 246]. Bisra et al. [247] 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. [248] 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.

E.2 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 agents' decisions is known as the Theory of Mind (ToM) [249, 250] in psychology. Based on this theory, Bayesian Theory of Mind (BToM) [251] provides a probabilistic framework to predict the inferences that people make about the mental states underlying other agents' actions [252]. Recent work, at the intersection of XAI and robotics, indicates that humans also attribute ToM to artificial agents that they observe or interact with [253, 254]. Guided by these user-centered results, several works at the intersection of XAI and robotics have utilized BToM to create a simulated user and then use the simulated user to generate helpful explanations. Towards this goal, Huang et al. [198] provide a greedy algorithm for selecting explanations that maximize the simulated user's knowledge of the agent's (a self-driving

	Studied Paper	Metric	Definition Source	Detail
Observed	[1]	Weight of Advice (WOA)	-	Degree to which the algorithmic suggestion influences the participant's estimate.
	[9, 11, 12, 15, 25, 26]	Agreement rate	-	Percentage of cases in which participants agree with the model. [12] defines the *appropriate trust*, *overtrust* and *undertrust*. [11] defines as *adherence*
Self-reported	[2, 3, 21]	Trust in Automation	[161]	On the 7-point Likert scale. [2] adapts the questions.
	[7, 23]	General trust in technology	[163]	On the 5-point Likert scale.
	[5]	Human-Computer Trust	[230]	On the 7-point Likert scale. [5] adapts the questions.
	[8]	Trust-TAM (Technology Acceptance Model)	[231]	On the 7-point Likert scale. [8] includes other self-designed questions.
	[22]	Trust in human-machine systems	[232]	On the 7-point Likert scale.
	[28]	Unified Theory of Acceptance and Use of Technology Model (UTAUT)	[233]	On the 5-point Likert scale.
	[10]	Human-Robot Collaborative Fluency Assessment	[125]	On the 7-point Likert scale
	[23]	Trusting beliefs and intentions	[234]	On the 7-point Likert scale.
	[6, 11, 13, 27, 29, 30] [16, 17, 18, 19, 24, 26]	Self-designed questionnaire	-	[6, 11, 17] are on the 7-point Likert scale. [13, 18, 19, 24, 27] are on the 5-point Likert scale. [16] rates from 0 to 100. [13, 16, 18, 19, 24, 29, 30] measure one-dimensional trust.
	[4]	Semi-structured interview	-	

TABLE 8: 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.

car in their domain) policy; and Lee et al. [255] 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 [185] 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 an 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.

E.3 Hybrid Teaching

Teaching strategies for the human-to-human setting have been widely studied and many categorizations exist [256, 257, 258]. 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 [255, 259, 260, 261, 262, 263]. 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 [264, 265, 266]. In XAI-related work, Qian and Unhelkar [185] 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 reduces the amount of time for a user to understand an agent's policy compared to direct and indirect teaching, and is more subjectively preferred by the participants. Building on this, future XAI systems can consider using hybrid teaching methods that (i) generate direct instructions to provide guidance to users' understanding of an AI system and (ii) provide methods to allow users to interact with the agent or model enabling active learning.

	Studied Paper	Metric	Definition Source	Detail
Workload	[3, 16, 21, 31, 66]	NASA TLX	[165]	
	[48]	Memory Performance	-	
Helpfulness	[13, 56, 65] [46, 48, 67]	Self-designed questionnaire	-	[13, 56, 65] are on 5-point Likert scale [46, 48, 67] are on 7-point Likert scale
	[68]	Rating	-	Rating from 1 to 5
	[45]	Comparison	-	Users select the most helpful method
Satisfaction	[6, 16, 18, 19, 29, 30, 70]	Self-designed questionnaire	-	[18, 19, 70] are on 5-point Likert scale [6, 29, 30] are on 7-point Likert scale [16] rates from 0 to 100
	[7, 69]	User experience of recommendation system	[163]	[7] adapts the questions on the 5-point Likert scale [69] adapts the questions on the 7-point Likert scale
	[1, 47]	Explanation Satisfaction Scale	[166]	[1] are on 5-point Likert scale [47] are on 6-point Likert scale
Undesired behavior detection	[71]	Number of identified bugs	-	Questions about bug identification and solutions
	[24]	Accuracy (percentage of correct answers) and time	-	Task is to determine the correctness of model answers
	[53]	Deviation between human's and model's predictions	-	Model's predictions are buggy and human's predictions should be different.
	[57, 72]	Accuracy (percentage of correct answers (bias detection))	-	Task is to identify (ir)relevant features
Ease of use and others	[24, 48, 65, 66, 81]	Self-designed questionnaire	-	[24, 65] are on 5-point Likert scale [48, 81] are on 7-point Likert scale
	[83]	AVAM and UEQ-S	[235, 236]	Autonomous Vehicle Acceptance Model Questionnaire (AVAM) [235] User Experience Questionnaire-Short (UEQ-S) [236] Both on the 7-point Likert scale
	[32]	Single Ease Question (SEQ)	[237]	On the 7-point Likert scale
	[71]	User Engagement Scale (UES)	[238]	On the 7-point Likert scale
	[37, 82]	System Causability Scale	[167]	On the 5-point Likert scale
	[3]	System Usability Scale	[239]	On the 5-point Likert scale
	[20, 85]	semi-structured interview	-	-

TABLE 9: 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.

		Other Evaluation Dimensions	
		Positive	Non-positive / Mixed
Trust		[2]: balanced training data, [17]: high model performance [23]: high quality of explanations [27]: high AI literacy [29]: interactivity [15]: model confidence	[2]: user expertise, insignificant [18]: personal characteristics, insignificant [30]: different topic modeling approaches, insignificant [5]: self-referential pronoun "I" in explanations, negative [22]: user technical literacy, insignificant
Understanding	Obj.	[32]: disentanglement of gen. model [22]: interactivity [45]: ExpO regularization of the model	[32, 53]: high dimensionality, negative [47]: contextualization, insignificant [13]: inductive vs. deductive explanations, insignificant [50]: different ML models, insignificant [52]: user expertise, insignificant [58]: instant feedback, insignificant [56]: timing of model errors, mixed
	Sub.	[32]: disentanglement of gen. model [22]: interactivity [44]: user expertise	[40]: model correctness, insignificant [36]: QuickSort, insignificant [35]: test of understanding, negative
Usability		[32]: significant difference in self-reported difficulty dependent on the generative model [7, 29]: interactivity [66]: Parallel Embeddings [27]: high AI literacy [73]: fair features are "current charges" and "criminal history"	[30]: different topic modeling approaches, insignificant [18]: personal characteristics, insignificant for satisfaction [56]: early encounters of system weaknesses lead to lower explanation usage [53]: clear model is *less* useful in debugging [73]: unfair features are "quality of school life" and "education & school behavior", etc.
Human-AI Collaboration Performance		[90]: low model complexity [13, 92]: showing model prediction	[10]: explanations are positive for novices' performance but negative for experts' [15, 43] Showing predictions, insignificant [15]: model confidence

TABLE 10: User study findings when using **other aspects** (other than the presence of explanation) 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.

		Evaluation Dimension: Explanations Effect comparison among **different explanations**
Trust		[9]: example-based explanations are positive in trust building [13]: deductive (rule-based) explanations > inductive (example-based) explanations in decision-making tasks, but contrary in proxy tasks [17]: different explanations positively affect different beliefs of trust [19]: proposed explanation interfaces (different visualizations), SCATTER > RANK and SCATTER > TUNER but insignificant [14] HEX-RL (theirs) > LSTM-attention (for RL agents)
Understanding	Obj.	[46]: Cues and Counterfactuals > Saliency (audio data) [48]: Sparse Lin. > COGAM > GAM [49]: MAME > SP-LIME [51]: CLUE > Sensitivity, Human CLUE, Random (for uncertainty) [52]: Natural images > synthetic (activation prediction) [57]: Counterfactuals (INN) = (proposed) Baseline Expl. > Concepts [55] Anchors > LIME
	Sub.	[33]: local+global explanation > local/global explanation [17]: example-based explanations (normative/comparative) improve the subj. understanding [40] LIME ≥ Composite, Prototypes and others [37]: closest and plausible counterfactuals, difference insignificant [33]: local+global explanation > local/global explanation [14] HEX-RL (theirs) > LSTM-attention (for RL agents) [44]: visual > textual explanations
Usability		[48]: sLM ≤ COGAM < GAM, insignificant for self-reported cognitive load [47]: contextualizing/exploration improve user's satisfaction, but no significant impact when interacting both factors [81]: diff. expl. (e.g. local expl., counterfactuals,...) [21]: GAM vs. SHAP, pos. for cognitive load [16]: diff. interfaces, pos. for cognitive load [46]: counterfactual+cues > saliency, pos. for helpfulness [68]: DEAML > EFM (feature-level expl.) > PAV ("people also viewed" expl.) for usefulness in RS [56]: Salient video segments > Confidence scores, Component combinations shown for helpfulness [13]: 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 [37]: closest and plausible counterfactuals, difference insignificant [69]: text explanation > visual explanations in user experience (e.g., satisfaction) [19]: proposed explanation interfaces (different visualizations), SCATTER > RANK and TUNER > SCATTER in satisfaction, RANK > SCATTER and TUNER > SCATTER in usefulness, but all insignificant [57]: Counterfactuals (INN) = (proposed) Baseline Expl. > Concepts in bias detection [72]: AReS (theirs) > AR-LIME [74]: sensitivity- and case-based explanations are rated as least fair when they expose a bias of the model [76]: acceptance of the gender-aware career recommender > gender-debiased [75]: significant preference for equalizing false positives over equalizing accuracy [27]: the amount of information positively relates with perceived fairness [2]: data-centric explanations that indicate balanced training data raise the fairness rating
Human-AI Collaboration Performance		[13]: both deductive (rule-based) explanations and inductive (example-based) explanations are positive, no significant difference

TABLE 11: User study findings when using model **explanations** as evaluation dimensions and comparing different explanation types on measured quantities.