

Towards Human-centered Design of Explainable Artificial Intelligence (XAI): A Survey of Empirical Studies

SHUAI MA, The Hong Kong University of Science and Technology, China

With the advances of AI research, AI has been increasingly adopted in numerous domains, ranging from low-stakes daily tasks such as movie recommendations to high-stakes tasks such as medicine, and criminal justice decision-making. Explainability is becoming an essential requirement for people to understand, trust and adopt AI applications.

Despite a vast collection of explainable AI (XAI) algorithms produced by the AI research community, successful examples of XAI are still relatively scarce in real-world AI applications. This can be due to the gap between what the XAI is designed for and how the XAI is actually perceived by end-users. As explainability is an inherently human-centered property, in recent years, the XAI field is starting to embrace human-centered approaches and increasingly realizing the importance of empirical studies of XAI design by involving human subjects.

To move a step towards a systematic review of empirical study for human-centered XAI design, in this survey, we first brief the technical landscape of commonly used XAI algorithms in existing empirical studies. Then we analyze the diverse stakeholders and needs-finding approaches. Next, we provide an overview of the design space explored in the current human-centered XAI design. Further, we summarize the evaluation metrics based on evaluation goals. Afterward, we analyze the common findings and pitfalls derived from existing studies. For each chapter, we provide a summary of current challenges and research opportunities. Finally, we conclude the survey with a framework for human-centered XAI design with empirical studies.

Author's Note: This manuscript was written in 2022 May, so the surveyed literature is not up-to-date. During the writing, I refereed a lot from Vivian Lai's paper [78] and Vera Liao's paper [89]. Since May 2022, many empirical studies on XAI have been published. Nevertheless, I hope this manuscript can serve as a starting point for interested readers.

Additional Key Words and Phrases: Explainable AI, Human-Centered Design, Empirical Study, Human-Centered AI

ACM Reference Format:

Shuai Ma. 2024. Towards Human-centered Design of Explainable Artificial Intelligence (XAI): A Survey of Empirical Studies. 1, 1 (October 2024), 36 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

Author's address: Shuai Ma, The Hong Kong University of Science and Technology, Hong Kong, China, shuai.ma@connect.ust.hk.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2024 Association for Computing Machinery.

Manuscript submitted to ACM

Manuscript submitted to ACM

CONTENTS

Abstract	1
Contents	2
1 Introduction	3
2 Commonly Adopted XAI Techniques in Empirical Studies	4
2.1 Global Explanation.	4
2.2 Local Explanation	6
2.3 Summary and Thinking	7
3 Different Stakeholders and Various Explainability Needs	8
3.1 User Group-Driven Explainability Needs	8
3.2 User Research-Driven Explainability Needs	8
3.3 Summary and Thinking	9
4 Design Space in Current Empirical Study	10
4.1 Explanation-related Design.	10
4.2 Model and Prediction-related Design.	12
4.3 Human-AI Collaboration Mode-related Design.	14
4.4 Summary and Thinking	15
5 Evaluation of XAI Design	18
5.1 Evaluation Related to Users' Perceptions of AI	18
5.2 Evaluation Related to Users' Interaction Experience.	20
5.3 Evaluation Related to Task	21
5.4 Summary and Thinking	22
6 Findings and Pitfalls Derived from Empirical Studies	23
6.1 Common findings	23
6.2 Pitfalls	25
6.3 Summary and Thinking	27
7 A framework for Human-Centered Design of XAI with Empirical Studies	28
7.1 Identify Design Goal	29
7.2 Design Explanatory Interface	29
7.3 Evaluate with Human Subjects	30
7.4 Summary	31
References	31

1 INTRODUCTION

With the advances of AI research, AI has been increasingly introduced into different tasks in people’s work and life, ranging from low-stakes daily tasks such as movie recommendations, virtual assistant agents, to high-stakes tasks such as medicine, criminal justice, etc [6, 36, 39, 68, 97, 137, 144]. When users interact with a system, users’ “mental model” (an understanding of how the system works) [106], plays a fundamental role for users to correctly and effectively interact with the system. However, many AI models are hard for users to understand due to their black-box nature [2]. This lack of transparency can lead to users’ inappropriate mental models of how the model works, which can cause other problems such as inappropriate trust and unexpected adoption of the intelligent system [22]. Besides, a lack of explainability of AI might cause failures in usability and moral crises, such as fairness, reliability, safety, accountability, etc [89]. Furthermore, some legal requirements have been proposed [2, 8], such as GDPR’s requirement that AI applications must provide people who are affected by automated systems with “meaningful information about the logic involved”.

To solve these problems, a lot of eXplainable Artificial Intelligence (XAI) algorithms have been introduced, where the AI model explains its reasoning process to the users [101]. Explanations can be any information that is beneficial for users to understand the AI model, ranging from logical information about the model, reasons for a prediction, to general information about the training data, input/output space, and more [8, 87, 89].

Many works in the XAI domain make technical contributions, such as developing new algorithms that increase the interpretability of the AI model as well as ensuring the model’s performance [2]. Meanwhile, the XAI community is increasingly realizing the importance of empirical studies of XAI design by involving human subjects. On the one hand, to design an effective and appropriate explanation, designers need to understand users’ specific needs in the target task based on user research [89]. On the other hand, the effectiveness of explanation is determined by the perception and reception of the explainees (receiver of an explanation, i.e., the human) [43] because they own different knowledge, expertise, background, role, needs, and goal for receiving the explanation [89].

The goals of XAI empirical studies are not only to evaluate the effectiveness of a specific XAI design, but also for researchers and designers to better understand how target users actually interact with the designed XAI to solve real-world tasks. With a comprehensive understanding of users’ perceptions and utilization of the explanatory interfaces, empirical studies can contribute on multiple grounds, including: (1) to guide new explainable techniques that provide more interpretable and effective explanations; (2) to provide insights for designers to design more effective explanation for designated human-AI interaction scenarios; (3) to understand how to integrate the knowledge of human’s cognitive processes and mental models, such as theories from psychology, cognitive science, social science, into human-centered XAI design.

Recently, a research community of human-centered XAI [44, 45, 89, 136] has emerged. However, there is a lack of systematic investigation of human-centered XAI design from the empirical study perspective. This hinders the emergence of systematic knowledge and joint research effort [78]. We recognize several challenges. First, existing XAI empirical studies are built on different XAI techniques and conducted with different stakeholders for different explainability needs. Without systematically reviewing the scope of these factors, we may not be able to form a whole picture of current XAI studies. Second, based on different design goals and research questions, different empirical studies tend to focus on various XAI designs. There is a lack of overview of the large design space, i.e., how different explanatory interactions are designed in existing empirical studies. Third, the evaluation method of empirical studies is varying for different research questions, evaluation goals, etc. This leads to the fact that the effects of XAI design

are measured by different evaluation methods in different studies, even though the design goals are similar. In other words, XAI design lacks a unified evaluation system. Lastly, existing empirical studies have identified a rich set of empirical findings which can help researchers and practitioners in the XAI domain understand the nuanced effects of different explanation designs. However, there lacks an organized summarization of these off-the-shelf findings, hindering knowledge sharing and transfer in this domain.

To move the first step towards a systematic review of empirical study for human-centered XAI design, we systematically investigate the current state of the field. We review works focusing on exploring human needs of XAI through user research, evaluating the effects of the proposed XAI design with empirical studies, and understanding humans’ perceptions and experiences when interacting with specific XAI designs, rather than works focusing on developing new XAI algorithms or developing a system without any empirical studies. The scope of this survey differentiates from either prior surveys on XAI that deviate from the focus of empirical study [44, 45, 89, 126] or empirical studies of human-AI interactions that do not focus on XAI [4, 125, 135]. To mitigate the four above-mentioned challenges, our survey focuses on analyzing four aspects of empirical study in these surveyed papers: the stakeholders and needs, the design space explored in existing work, the evaluation metrics of different design goals, and the common findings and pitfalls from existing empirical studies.

The remainder of this survey is structured as follows. We first start with a brief overview of XAI techniques commonly adopted in existing empirical studies. Then we analyze the diverse stakeholders and need-findings approaches. Next, we show an overview of the current explored XAI design space. Further, we summarize the evaluation metrics based on the evaluation goals. Afterward, we analyze the common findings and pitfalls derived from existing studies. Finally, we conclude the survey with a framework for designing human-centered XAI with empirical studies.

2 COMMONLY ADOPTED XAI TECHNIQUES IN EMPIRICAL STUDIES

There are some relevant terms with “explainability”, such as “interpretability”, “intelligibility”, “transparency” [101]. In this survey, we adopt the most commonly used term “explainability” which shares a goal to make AI understandable to users. Recent papers have surveyed this field from a technical perspective and converged on several important dimensions to classify XAI algorithms [2, 8, 10, 88]. One categorization is based on the intrinsic transparency of the AI models, which can be divided into directly explainable models (such as rule-based models, linear model, decision trees) and opaque models (such as deep neural networks), and the latter often require using additional algorithms to generate post-hoc explanations [57, 92]. Another dimension widely used in XAI classification is the scope of explanation. XAI techniques can be categorized into *global* explanation of the overall logic of the model, and *local* explanation of a specific prediction [89]. In this chapter, we also use this dimension to categorize and conclude common explanation methods adopted in existing empirical studies for XAI, and describe each briefly. Note that we will treat all explanatory approaches as XAI methods as long as they can benefit users in understanding the AI. So, apart from model inner working-related explanation and feature-related explanation, we also include model-related factors into our XAI methods coverage, such as training data-related explanation, uncertainty-related explanation, performance-related explanation, etc.

2.1 Global Explanation.

Stakeholders of AI applications often need to understand the underlying logic of an AI model to form an accurate mental model. “Global” explanation about the model can include global feature importance, global example-based explanations, presentation of simple models, model’s overall performance, model documentation, information about the training data, etc.

Table 1. XAI Techniques Category Based on Explanation Scope.

Scope	Category	Method and Example
Global	Global feature importance	Permutation-based [49, 138], coefficients [40]
	Global example-based explanations	Model tutorial [79], prototypes [25, 48, 104]
	Presentation of simple models	Decision trees [121], linear regression [111], decision sets [81], logistic regression [121], one-layer MLP [121]
	Model performance	Accuracy [60, 79, 80, 143, 145], false positive rates [60]
	Model documentation	Overview of the model or algorithm [70, 73, 74, 85, 113], model prediction distribution [134]
	Information about training data	Input features or information the model considers [40, 60, 111, 148], aggregate statistics (e.g., demographic) [16, 40], full training “data explanation” [5]
	Model uncertainty	Classification confidence (or probability) [9, 13, 20, 22, 48, 59, 83, 91, 148]
Local	Example-based methods	Nearest neighbor or similar training instances [19, 25, 26, 40, 61, 80, 133, 138]
	Local feature importance	Coefficients [31, 40, 52, 55, 79, 80, 93, 111], attention [28, 29, 79], gradient-based [29, 69, 105], propagation-based (LRP [3]), perturbation-based (LIME [3, 61, 105], SHAP [140, 148])
	Rule-based explanations	Decision sets [74], tree-based explanation [76], anchors [117]
	Counterfactual/Contrastive explanations	Contrastive or sensitive features [40, 96], counterfactual examples [121, 138]

2.1.1 Global feature importance. Global feature importance quantifies the overall importance of each feature used to get the model’s decisions. Some models can directly produce feature importance, such as coefficients in linear or logistic regression models [40] and shape function of GAMs [1]. Another way to get feature importance is from the post-hoc model, such as permutation importance [49, 138]. Note that feature-importance methods can also be used for local explanation.

2.1.2 Global example-based explanations. Example-based explanations provide data examples for users to understand how the model works. One way is to select some instances from the training set that can provide insights to the user. For example, Lai et al. [79] select examples from the training set as a tutorial. Another common approach is that for a given prediction class, selecting one or a set of training instances that are representative and have the same class labels [104]. The example-based explanation can also be used to locally explain a prediction. The main difference is that the representative instances are often provided in the onboarding stage for the global explanation.

2.1.3 Presentation of simple models. For a complex black-box model, it is hard for users to understand its complex internals. Usually, post-hoc XAI methods are utilized to train a simple directly interpretable model such as a decision tree, rule set, with the same training data to provide an approximate overview of how the complex model behaves [37, 89, 130]. For a simple model, we can directly present the model internals to users.

2.1.4 Model performance. Model performance can provide the basic information of how well the model works in general to users. In the empirical studies with classification tasks, model performance is mainly presented in the form of accuracy [60, 79, 80, 143, 145]. These works typically explore the effects of whether to show or how to show the model performance on users’ perception of and decision making with the model [80, 145]. Note that accuracy is usually estimated on the validation set, and the model’s actual performance can be different from the estimated performance because the model has to perform prediction on unseen data.

2.1.5 Model documentation. Some literature designs model documentation to provide the meta information of a model. This meta information includes the characteristics of the model, how and for what purpose the model is developed, which is verified to be critical to AI transparency [7, 78, 100]. For example, some “About Me” page Model cards have developed [7, 100]. Kocielnik et al. [70] propose a meeting scheduling assistant that displays a description of how it works.

2.1.6 Information about training data. As the training data plays a critical role in the model development, providing the information about the training data, such as data distribution and feature set, is verified to be able to help users better understand the model [40, 60, 111, 148]. For example, Zhang et al. [148] investigate whether the additional knowledge can affect users’ trust in a simulated income prediction task, where the participants are made aware of whether or not the model has considered “marital status” as a feature.

2.2 Local Explanation

Local explanations are usually used to explain how a specific prediction is made by the model. The explanation information of a local prediction includes local feature importance, example-based explanations, rule-based explanations, counterfactual explanations, and model’s uncertainty for the prediction, etc.

2.2.1 Model uncertainty. Although a model’s uncertainty or confidence for a prediction is not a direct explanation for how and why the prediction is made, it can bring valuable information about how confident the model is for the prediction. Then the users can decide whether to rely on the model based on the uncertainty [53]. The uncertainty or confidence is usually calculated as the probability of the predicted label in a classification task, ranging from 0 to 1 (or scaled to 0 to 100). A lot of empirical studies have investigated the effects of showing model uncertainty on users [9, 13, 20, 22, 48, 148].

2.2.2 Example-based methods. Example-based methods use similar examples to the current instance to support case-based reasoning. Generally, for a target instance to explain, the model will research similar examples from the training data which have the same label/class as the target instance. The similar examples are usually found by some similarity/distance metrics in the embedding space or feature space. The example-based explanation had been widely adopted in many empirical studies [19, 25, 26, 40, 80, 133, 138].

2.2.3 Local feature importance. The local feature importance explanation will calculate the importance of each feature to the current prediction. For example, when predicting income, some features play a more important role to the prediction (e.g., the occupation) while others might be less important (e.g., the weight of a person). Generally, the local feature importance can be obtained in two ways, the built-in method and the post-hoc methods.

Built-in methods can be mainly categorized into the coefficient-based method and attention-based method. First, some models can generate coefficients as feature importance. For instance, the coefficient of a linear regression model can be seen as a direct measure of feature importance. Coefficient-based feature importance has been adopted in many empirical studies [31, 40, 52, 55, 79, 80, 93, 111]. Second, the attention-based method is often used for explaining deep learning models, such as attention in an NLP model and saliency region in a computer vision model. This method has also been verified in some works [28, 29, 79]. For example, Lai et al. [79] compare the attention-based explanation with a LIME-based explanation.

Post-hoc methods usually train a separate model for the original model, often used for black-box models. And the feature importance is generated from the new model. Post-hoc methods can be categorized into gradient-based

[29, 69, 105], perturbation-based [3, 32, 61, 148] and propagation-based [3]. From the existing studies, we find that LIME [3, 61] and SHAP [140, 148] are two widely used methods.

2.2.4 Rule-based explanations. Rule-based methods, such as a set of “if-then” explanations can be easily understood by humans. Similar to the feature importance explanation, the rule can be directly got from simple models, such as decision trees [76], decision sets [74]. Also, the rule can be generated from post-hoc methods, such as anchors [117]. Compared to example-based and feature importance-based explanations, the rule-based explanation is less investigated in existing empirical studies.

2.2.5 Counterfactual/Contrastive Explanations. Counterfactual explanations are widely used when users would like to know how the prediction will change if the current input changes. If a user gets a prediction that is not her expected, she might be interested in figuring out “how to change to get a different prediction” or “why not a different prediction”. For example, a loan declined user wants to know how to get approved. In some surveys [88, 89], counterfactual methods are differentiated from local explanation because it is not the direct explanation for a prediction. However, in this paper, we categorize it into the local scope as it is often desired after a user sees a specific prediction.

In the research literature, contrastive explanations are often conflated with counterfactual explanations [147]. Although similar, they have differences. Generally, contrastive explanations are used to answer “Why Not” questions, and counterfactual explanations focus on answering “How To Be That” questions [90].

2.3 Summary and Thinking

There have been a rich amount of XAI techniques which offer opportunities to design explanation for both simple models and complex black-box models. Categorized by the explanation scope, these techniques can be divided into global methods and local methods. XAI designers can select appropriate XAI techniques based on the type of AI models and the specific explanation goals. We also find some gaps.

2.3.1 Challenges. Limited AI types and task types applied in empirical studies. From the surveyed papers, we find that most studies use simple models for interpretability. While for complex models, such as DNN, CNN, GNN, and RL, there is still a lack of empirical studies for non-expert users, although there are corresponding explanation technologies and some visual interaction systems have been developed to help model developers understand and debug models. This needs to solve several problems, first of all, although complex models can be approximated as simple models or explained by posthoc methods, the faithfulness issue should be considered. Second, the complex information brought by the explanation of complex models requires more expertise from ordinary users. Therefore, it is necessary to seek a balance between the accuracy and comprehensiveness of the explanation and the ease of interpretation.

Another trend is that in most of the current empirical studies, simplified experimental tasks are selected, usually a binary classification task is used as the testbed. However, tasks in the real world are not only binary classification, such as multi-classification or continuous prediction tasks (regression tasks). Although there are several works that choose regression tasks, such as apartment price prediction [111], research in this area is relatively rare.

2.3.2 Future research opportunities. Informing new XAI techniques by human-centered research. One way that human-centered research can contribute is to offer more implications to the technical development of XAI. One representative example is RexNet published at CHI2022 [147], which generates relatable explanations inspired by humans’ perceptual process from cognitive psychology. There are rich cognition, psychology, and social theories under exploration in XAI algorithm design.

Investigating the effects of more XAI algorithms in more diverse tasks. In the future, explanations for other types of AI should be explored in empirical studies, such as Reinforcement Learning models. And HCI researchers can pay more attention to regression tasks which are under explored.

3 DIFFERENT STAKEHOLDERS AND VARIOUS EXPLAINABILITY NEEDS

Different stakeholders of AI application have different needs for explanation [87, 89, 101]. To understand users' explainability needs, some papers categorize users into different groups, while another lines of papers adopt human-centered user research to derive user needs.

3.1 User Group-Driven Explainability Needs

There are mainly two types of user group categorization methods. One is based on users' expertise and knowledge, and the other is based on users' role when interacting with AI applications.

3.1.1 Expertise-Based User Group Categorization and Corresponding Explainability Needs. User expertise/knowledge is one commonly recognized characterization. Literature has shown that the XAI design goals can be different for different levels of expertise. For example, Mohseni et al. [101] categorize users into "AI experts", "data experts", and "AI novices". The design goals for AI experts can include model interpretability and model debugging. The design goals include model visualization and inspection, and model tuning and selection for data expert (also called domain expert in other papers, who is an expert in the AI application domain but has little knowledge of AI). And the design goals for AI novices can include AI transparency, user trust, bias mitigation, and privacy awareness. Recently, Suresh et al. [128] introduce a framework with a more granular characterization for the stakeholders based on their knowledge and objectives. From the expertise perspective, they characterize stakeholders' knowledge (i.e., formal, instrumental, and personal knowledge) and the context to which the knowledge belongs (i.e., machine learning, the data domain, and the milieu). From the objective perspective, they identify three-level of goals, from long-term goals to shorter-term objectives to immediate goals.

3.1.2 Role-Based User Group Categorization and Corresponding Explainability Needs. Another commonly recognized characterization of users is based on users' functional roles. For example, Arrieta et al. [8], Hind et al. [62] and Preece et al. [112] summarize several common user groups and map the corresponding needs, including (1) Model developers whose needs are to debug or improve a model; (2) Business owners or administrators whose needs are to assess an AI application's regulatory compliance and capability; (3) Decision-makers whose needs are to make informed decisions with appropriate trust; (4) Impacted groups, whose needs are to seek recourse or contest the AI; (5) Regulatory bodies whose needs are to audit for legal or ethical compliance such as fairness, safety, privacy, etc. Based on such categorization, the role a person plays during the human-AI interaction can determine their explainability needs.

3.2 User Research-Driven Explainability Needs

In the XAI domain, researchers have increasingly realized that a good XAI design needs to understand the actual needs of users. One of the most direct and effective methods might be to obtain first-hand user needs based on user research methods such as interviews, surveys, formative studies, participatory design, etc. For example, for the explanatory interface design of an AI-assisted admission prediction system, Cheng et al. [31] organize a workshop inviting several types of stakeholders, including algorithm experts, UI designers, prospective students interested in applying to the university, current graduate students, faculty members to join the workshop. For a COVID-19 symptom checker, Tsai

et al. [133] conduct a first user study to investigate what kinds of explanations users really need. To help clinical decision-making, Cheng et al. [30] develop VBridge. To figure out users' needs for explainable AI, they conduct a pilot study with six clinicians to understand their expected functions and explanation methods.

One representative work in recent years is the question-driven explainability needs categorization proposed by Liao et al. [87, 88]. In a work called questioning the AI [87], they propose to identify users' explainability needs by figuring out what questions users want to ask. They contribute an *XAI Question Bank*, with more than 50 detailed user questions organized in 9 categories, including *How*, *Why*, *Why Not*, *How to be That*, *How to Still Be This*, *What if*, *Performance*, *Data*, *Output*. Each category of questions is related to some types of explanations. This XAI question bank can be used as a tool to identify applicable questions in user research. In a follow-up work [88], they propose a question-driven user-centered design method. First, designers can identify questions users might ask through user research. Then, designers can choose XAI techniques and iteratively design the interface based on questions. It is worth mentioning that they map the XAI technique space (as well as corresponding open-source XAI toolkits) with the user question categories, which is very helpful to select specific XAI methods for a specific user question to solve.

3.3 Summary and Thinking

There is a problem with the *User Group-Driven* explainability needs acquisition because beyond users' expertise and knowledge, a large scope of user characteristics can affect how an explanation works for a specific user, such as users' locus of control [119], need for cognition [24], visual literacy [1, 18], etc. Due to the diversity of users, there are always complex intersections between different user groups, and even users belonging to the same group are still different in many other perspectives, and these differences can affect their explainability needs and the effects of XAI on them. We believe that the *User Group-Driven* method can play a guiding role in the design of XAI, and the two methods should be combined. Specifically, *User Group-Driven* method can be used as a benchmark to guide the direction of user research, such as the questions design in a pilot study. At the same time, in order not to be limited by the division of fixed user groups, when designing XAI in a new scenario, it is necessary to make full use of *User Research-Driven* method to obtain specific user needs, so as to better match it to specific task scenarios. There are already some good practices. For example, Cai et al. [27] and Tonekaboni et al. [132] follow this approach of role-based needs-finding as well by interviewing clinicians. However, we identify some gaps.

3.3.1 Challenges. Lack of actual user needs driven research The empirical studies we survey can be mainly divided into two categories, one is driven by research questions, and the other is driven by user needs. The former usually first raises several research questions which can be composed of independent variables (IVs) and dependent variables (DVs). Generally speaking, IVs are the design method of different XAI interfaces or interaction patterns (see Section 3), and DVs are specific research goals (see Section 4). After identifying the research questions, these studies will design a research prototype interface for a simulated task with some open dataset, and then recruit subjects to participate in the experiment. These research question-driven papers can provide empirical value to the field and can help guide the development of new XAI technologies and the design of XAI interfaces. However, we find that compared with research question-driven studies, user needs-driven studies are rare. User needs-driven studies first obtain the real needs of users through formative studies, interviews, or participatory design methods, to guide the design of the XAI interface. After the design, they evaluate the XAI design via user studies.

Lack of user adaptive XAI design In XAI design, a recognized concept is that the design of XAI should be user-centric and the difference between users directly affects the effects of XAI. In other words, what effect XAI can

achieve is not determined by the XAI designer, but depends on how the user receives and perceives the explanation. Based on this, some works classify the types of users based on their roles in AI applications, or based on their expertise and knowledge background, and summarize different user needs and design goals for different user groups. However, it is not enough to design user group-specific XAI, because even two users who belong to the AI-expert group may be different in many other aspects. Therefore, user-adaptive XAI should be developed.

3.3.2 Future research opportunities. Complementing research question-driven studies with actual user needs-driven studies. When conditions permit, researchers can focus more on real users in actual scenarios, and obtain their interpretation needs through user research methods. Although many AI applications have been deployed in people’s work and life, they lack effective explanation functions. Researchers can use existing knowledge in the XAI field to solve real-world explanation problems.

Building adaptive XAI based on user modeling methods. Instead of “one-size-fits-all” solutions, user-adaptive XAI is expected to adapt the explanations according to users’ dynamic information needs, which requires understanding and modeling users in multiple dimensions, including but not limited to the users’ ability to perform tasks, AI literacy, domain knowledge, need for cognition, visual literacy, and even the users’ inherent impression of AI. Further, the user model needs to be dynamically updated during the interaction.

4 DESIGN SPACE IN CURRENT EMPIRICAL STUDY

In this section, we present an overview of how different aspects and factors are designed or manipulated in existing XAI empirical studies. We believe that this part is critical for new practitioners as this can help them draw a big picture of the current developed design space in this area. The XAI design space is not only for designing different types of explanation methods, but also for other interaction-related factors correlated, such as interface design, and cognitive process design, etc. Note that to make the survey more focused and informative, we focus on empirical studies of XAI, excluding papers that mainly design an explainable system without empirical studies. Based on the surveyed papers, we categorize the design space into Explanation-related design, Model and prediction-related design, and Human-AI collaboration mode-related design.

4.1 Explanation-related Design.

Most surveyed papers focus on the study of explanation-related design, varying from the type, and modality to the interactivity and complexity. Figure 1 shows some examples of explanation-based design.

4.1.1 D1: W/, W/o, or Random Explanation. In current XAI empirical studies, a commonly used and intuitive method to investigate the effects of the proposed explanation design is to compare with a “no-explanation” condition. Note that the “no-explanation” condition in these papers may not be the only baseline, and there might be other baselines to compare with. A great number of works compare one type of explanation with baselines that are without explanations [13, 31, 47, 111, 118, 120, 138, 147, 148]. However, some research has found that even a meaningless explanation or randomly generated explanation can foster users’ positive perceptions [43, 80, 93]. For example, Ehsan et al. [43] find that even a meaningless numeric “explanation” can lead to over-trust of both AI experts and non-experts. Therefore, to counteract the users’ positive attitude toward explanations, some studies use non-informative explanations or random explanations as baselines [43, 43, 80, 93].

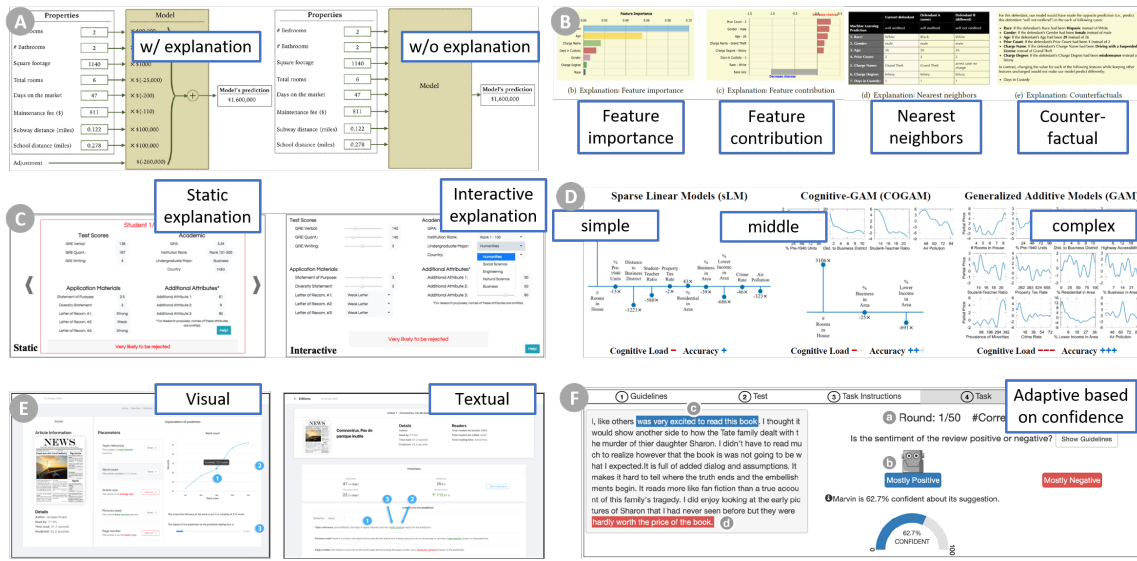


Fig. 1. Examples of explanation-related design. (A) w/ and w/o explanation [111]. (B) Explanation types [138]. (C) Explanation interactivity [31]. (D) Granularity/complexity of explanation [1]. (E) Modality of explanation [129]. (F) Explanation adaptability [13].

4.1.2 *D2: Explanation Type.* As mentioned in Sec. 2, a large number of explanation types has been proposed. To investigate the actual effects of these techniques on users, HCI researchers have designed a lot of explanation types and proposed empirical studies to investigate and compare different types of explanations in targeted tasks [40, 79, 103, 133, 138, 143, 147]. For example, Wang et al. [138] compare four types of common explanations used in literature, including feature importance, feature contribution, nearest neighbors, counterfactuals, and a controlled no-explanation. Recently, Zhang et al. [147] propose a model, RexNet that can generate three types of explanations, Contrastive Saliency, Counterfactual Synthetic, and Contrastive Cues explanations, and compare the effects of different explanation types. Based on a pilot study, Tsai et al. [133] design three explanation forms for an AI-based COVID-19 symptom checker, including Rationale-based Explanation, Feature-based Explanation, and Example-based Explanation.

4.1.3 *D3: Explanation Interactivity.* Interactive explanation is often seen in human-human interaction, which is a social nature of explanation [98]. Some social science works have argued that explanations should be interactive [94, 98].

Interactive explanation means that users can interact with the AI’s explanation through an interface, such as changing the attribute values of an instance [31] or creating *What-If* explanations [136], to inspect the updated prediction. The exploration during the interaction can help users understand how the model works. In the context of XAI, Cai et al. [25] present a interactive system that allows trial-and-error to explore how an image recognition algorithm works. For a graduate admission prediction task, Cheng et al. [31] compare the effects of a static explanation and an interactive explanation where users can modify a student’s profile to see the changes in prediction. And for human-AI decision making, Liu et al. [93] explore the effects of interactive explanations in three prediction tasks. To help data scientists understand their models, Krause et al. [72] design and implementation of an interactive visual analytics system, Prospector, providing interactive partial dependence diagnostics for users to understand how features affect the prediction.

4.1.4 D4: Explanation Complexity/Granularity. The complexity or granularity of an explanation represents the complexity or the amount of detailed information contained in the explanation. In recent studies, researchers begin to pay attention to this dimension. For instance, Lage et al. [77] conduct a user study to investigate how explanation complexity affects participants' comprehension and performance. Poursabzi-Sangdeh et al. [111] manipulate the complexity of a model to alter the interpretability of a model: the number of features used in the model and the transparency of the model. Abdul et al. [1] proposed COGAM, a model that can adjust the complexity of an explanation based on users' cognitive load. For a sentiment classification task, Springer et al. [127] compare different explanations with different levels of complexity. Mishra et al. [99] explore granularity (of data features) and context (of data instances) as dimensions, and investigate the effects of granularity, context, simplification affect user understanding and confidence in ML models.

4.1.5 D5: Explanation Modality. The modality in which explanations are presented is an important characterization affecting how users will perceive it [101, 129]. Explanations could be presented in multi-modality, such as textual [114, 133, 141], verbal [64], visual in the form of graphs (e.g. saliency maps [2], scatterplot diagrams [72], bar chart [148], attention map [79], line charts [136]), etc. Some researchers have begun to be interested in explained modalities [34, 108]. For example, in a recommender system, Kouki et al. [71] find that textual explanations are more persuasive than visual explanations. Hohman et al. [64] combine visual explanations with textual explanations to vary the complexity. Szymanski et al. [129] propose three modalities of explanation, textual explanation, visual explanation, and hybrid explanation. Through a user study on video game players, Robertson et al. [118] find that when users' attention resources are limited, presenting explanation interventions via different modalities, like audio, interactivity, and text, can aid real-time comprehension. Due to the limited number of research on the modality of explanation, more efforts are needed to explore how and when to combine multi-modal explanations for a unified goal.

4.1.6 D6: Explanation Adaptability. Besides the traditional pre-defined explanations, from the surveyed papers published in recent two years, we also find several works proposing adaptive explanations, which show a promising direction for better dynamic explanations to cater to users' various cognitive abilities and needs. For example, Bansal et al. [13] introduce an adaptive explanation. It tries to reduce human trust when the AI has low confidence: it only explains the predicted class (Top-1-explanation) when the AI is confident, but also explains the alternative (Top-2-explanation) otherwise. Rastogi et al. [114] propose a confidence-based adaptive time allocation strategy to allocate a different amount of time for users to perceive and understand an explanation. Abdul et al. [1] propose COGAM, an adaptive explanation complexity adjustment method by calibrating visual cognitive chunks with users' potential cognitive load, and the COGAM can achieve the trade-off between cognitive load and accuracy.

4.2 Model and Prediction-related Design.

In addition to the design directly related to explanations, there is some information that can also play a role in helping people understand AI, including information related to models and predictions. Figure 2 shows the examples of model and prediction-related design.

4.2.1 D7: W/ and W/o Tutorial/Training. Some studies have found that only presenting global or local explanations during the task process is not effective enough for users to fully understand the AI [7, 79, 100]. They pay attention to the tutorial process before the task. For example, instead of directly providing real-time assistance to people, Lai et al.

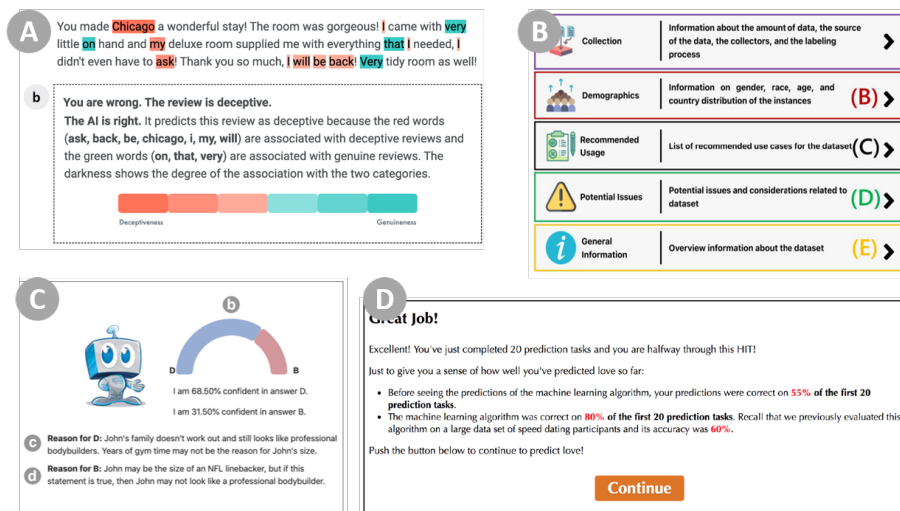


Fig. 2. Model and prediction-related design. (A) W/ and W/o Tutorial/Training [79]. (B) Explanation of Training Data [5]. (C) W/ and W/o Confidence/Uncertainty [13]. (D) Model performance [145].

[79] focus on the training stage, and propose tutorials to help users understand the patterns embedded in a model and the nature of a task. Particularly, they propose two kinds of tutorials, guideline-based and example-driven tutorials.

4.2.2 D8: Explanation of Training Data. Some research finds that providing the information of training data can facilitate users' understanding of the model, such as the input features used or data distribution [40, 60, 74, 111, 148]. For example, in an income prediction task, Zhang et al. [148] investigate the effect of humans are made aware of whether or not the model considers "marital status" as a feature. Liu et al. [93] investigate the effects of telling users the data distribution on human-AI decision making. To promote transparency, Anik et al. [5] propose data-centric explanations to explain a series of information of training data, including how the data was collected, demographics, recommended usage, potential issues, etc.

4.2.3 D9: W/ and W/o Confidence/Uncertainty. Controlling whether showing a model's confidence or uncertainty is one of the most common designs in human-AI collaborative decision-making studies. For example, Zhang et al. [148] investigate the effects of confidence scores on users' trust, and accuracy of AI-assisted predictions. Rastogi et al. [114] propose a confidence-based time allocation strategy for AI-assisted decision-making, and find that when the AI model has low confidence and is incorrect, their proposed confidence-based time allocation strategy can effectively mitigate users' cognitive biases and improve the collaborative performance. Bansal et al. [13] compare different explanations with a confidence-based condition and find that confidence scores can potentially help people form a good mental model of AI's error boundaries.

4.2.4 D10: Model performance. Another important piece of information is the model's performance in the training dataset, which is shown to be able to affect users' trust. In classification tasks, model performance is mainly presented in the form of accuracy [60, 79, 80, 143, 145]. There have been some studies exploring how showing model accuracy to users affects their perception of and task performance. For example, Lai et al. [79] investigate whether the model's accuracy improves users' performance in decision-making tasks. And, Yin et al. [145] study the effect of accuracy

on human trust in ML models. Model performance can also be measured by precision and recall. For example, in a recidivism prediction task, Harrison et al. [60] show that presenting false positive rates can help people judge the fairness of the model. Considering that model performance estimated in the training data can be inconsistent with the real performance, Yin et al. [145] investigate the effect of communicated accuracy and experienced accuracy on users.

4.3 Human-AI Collaboration Mode-related Design.

Figure 3 shows the examples of human-AI collaboration mode-related design.

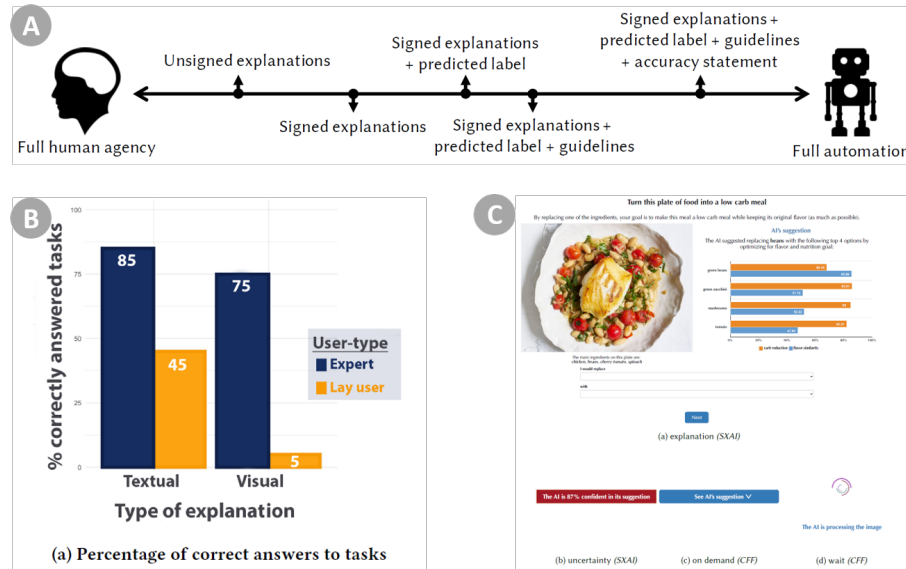


Fig. 3. Human-AI collaboration mode-related design. (A) AI Agency (Degree of Explanation) [80]. (B) Expertise of Explanation Users [129]. (C) Cognitive Bias Mitigating [20].

4.3.1 D11: AI Agency (Degree of Explanation). AI can have different levels of agency in the human-AI interaction process. In the context of XAI, the agency can be determined by the degree of explanation. For example, a model only with an explanation is a low agency, while a model with prediction, confidence, and explanation is high agency. Researchers have compared different levels of machine agency. For example, Buçinca et al. [20] examine the effects of on-demand explanation that is receiving predictions only on demand. For a deception detection task, Lai et al. [80] compare different AI agencies, including 1) Control-human only, 2) Feature-based explanations, 3) Example-based explanations, 4) Predicted label without accuracy, 5) Predicted label with accuracy, 6) Combinations. Levy et al. [86] compare two clinical note annotation systems with different levels of AI agency, one only suggests annotation labels after users choose text spans to be labeled, and another performs both span and label suggestions.

4.3.2 D12: Expertise of Explanation Users. It has been well recognized in the XAI area that user expertise is a key aspect to be considered when designing an explanatory interface [8, 21, 126]. It is an open challenge in XAI to tailor the explanations to the expertise of the end-user [58].

In the empirical studies that focus on user expertise as an independent variable, Szymanski et al. [129] investigate how different levels of expertise influence the understanding of explanations. Wang et al. [138] investigate the effects of different explanations on two tasks where participants have more expertise in one task but not in the other. Anik et al. [5] test their proposed explanation on users with different expertise, i.e., expert, intermediate, beginner. In a graduate admission prediction task, Cheng et al. [31] explore how the effectiveness of the explanation interface can be influenced by users' personal characteristics (i.e. education level and technical literacy). Besides, Schaffer et al. [120] explore the difference between participants' measured expertise and their self-reported familiarity in an AI-assisted binary decision-making game. In addition, Ehsan et al. [43] conduct a mixed-methods study to investigate how users with different AI backgrounds perceive different types of AI explanations.

4.3.3 D13: Cognitive Bias Mitigating. How humans perceive the explanation is at the core of whether and how an explanation works for them [114]. However, human cognition is inclined to natural flaws, like cognitive biases [147]. Recently, a number of empirical studies have focused on designing cognitive de-bias methods to better communicate the explanation to the user [20, 114]. Evidence shows that providing interventions to change users' cognitive processes can significantly affect how they process the explanation information to form mental models of AI and make decisions [78].

One method to mitigate the cognitive bias is designing a more appropriate workflow. The typical paradigm of AI-assisted interaction is to have the AI provide its suggestion, then the user can choose to take the AI's advice or follow her own decision. Some studies explore the effects of having users make their own predictions before being shown the model output [20, 56, 95, 111, 138, 145, 148]. Such designs are verified to be able to force people to think more analytically rather than merely relying on the model's predictions. Existing work also explores the impact of different workflow designs on users' mental models, such as including a training phase before the task [29, 79, 111, 148], showing users their own or models' decision correctness [12, 13, 29, 56, 143, 146]. Focusing on the anchoring bias, Nourani et al. [107] investigate how the order of observing system weaknesses and strengths can affect the user's mental model, task performance, and reliance.

Another method to mitigate the cognitive bias is modifying the system response time [20, 109]. For example, Buçinca et al. [20] propose and explore the effects of three kinds of cognitive forcing functions. One of them is slowing down the process that the AI offers its prediction. Rastogi et al. [114] propose an adaptive time-allocation method and find that allocating more time can alleviate anchoring bias, especially in the condition that AI makes a mistake.

4.4 Summary and Thinking

Existing human-centered XAI research has explored a wide design space, which fall into explanation-related design, mode and prediction-related design, human-AI collaboration model-related design. An overview of the design space can help practitioners make better design choices.

4.4.1 Challenges. Lack of XAI design for the actual scene Currently, most XAI empirical studies simulate real task scenarios by designing an experimental interface. However, there are several differences between this experimental process and the real task scene. The first difference is in the task context and the process. In the empirical study, a simplified task process is used. After being told the experimental task, the user will directly enter the decision-making trials. However, in the actual task, other steps that seem not directly related to the decision-making are likely to affect the user's perception of XAI. The second difference is between experimental data and real data. Due to some limitations, most experiments are now based on public data sets. However, the composition of these public data sets could be different from real-world data. The third difference is the end user. Many studies now choose to use crowdsourcing

Table 2. Summary of a subset of the surveyed literature organized by the two dimensions, design Space (independent variable) and evaluation goal (dependent variable)

Work	Design Space										Evaluation Goal											
	Explanation-related Design						Model and Prediction-related Design				Collaboration Mode-related Design			Evaluation related to users' perceptions of AI				Evaluation related to users' interaction experience			Evaluation related to the task	
	D1: W/, W/o, or Random Explanation	D2: Explanation Type	D3: Explanation Interactivity	D4: Explanation Complexity/Granularity	D5: Explanation Modality	D6: Explanation Adaptability	D7: W/ and W/o Tutorial/Training	D8: Explanation of Training Data	D9: W/ and W/o Confidence/Uncertainty	D10: Model performance	D11: AI Agency (Degree of Explanation)	D12: Expertise of Explanation Users	D13: Cognitive Bias Mitigating	G1: Understanding/Mental Model	G2: Trust/Reliance	G3: Trust Calibration	G4: Fairness Judgment	G5: Satisfaction/Perception of the system	G6: Cognitive Load/Mental Demand	G7: Engagement	G8: Task Performance	G9: Time Spent/Task Efficiency
Wilkinson et al. 2021 [141]	✓												✓	✓								
Szymanski et al. 2021 [129]					✓						✓		✓					✓				
Chromik et al. 2021 [32]							✓						✓									
Wang et al. 2021 [138]	✓	✓									✓		✓		✓							
Buçinca et al. 2020 [19]	✓	✓											✓	✓				✓				✓
Zhang et al. 2020 [148]	✓							✓							✓							✓
Lai et al. 2019 [80]										✓				✓								✓
Fan et al. 2022 [47]	✓											✓	✓	✓				✓	✓			✓
Rastogi et al. 2022 [114]						✓			✓			✓										✓
Buçinca et al. 2021 [20]								✓				✓			✓			✓	✓			✓
Zhang et al. 2022 [147]	✓	✓											✓					✓				✓
Sangdeh et al. 2021 [111]	✓			✓									✓		✓							✓
Bansal et al. 2021 [13]	✓					✓		✓										✓				✓
Tsai et al. 2021 [133]		✓											✓	✓				✓	✓			
Anik et al. 2021 [5]								✓			✓			✓		✓		✓	✓			
Lai et al. 2020 [79]		✓					✓		✓	✓				✓		✓		✓				✓
Abdul et al. 2020 [1]				✓		✓							✓					✓				
Cheng et al. 2019 [31]	✓		✓								✓		✓	✓								✓
Robertson et al. 2021 [118]	✓	✓	✓		✓								✓					✓				
Narkar et al. 2021 [103]		✓																✓				
Nourani et al. 2021 [107]												✓	✓	✓								✓
Yang et al. 2020 [143]		✓													✓							
Springer et al. 2019 [127]				✓									✓	✓								
Schaffer et al. 2019 [120]	✓									✓				✓								
Dodge et al. 2019 [40]		✓														✓						
Liu et al. 2021 [93]			✓				✓											✓				✓
Mishra et al. 2021 [99]				✓									✓									
Ehsan et al. 2021 [43]											✓		✓	✓				✓				
Yin et al. 2019 [145]									✓					✓				✓				

experimental methods to obtain sufficient data. However, it can lead to some problems, such as fidelity, especially when the task setting requires a certain professional background. With these three differences, XAI design can be evaluated in an inappropriate manner and the findings may not be able to generalize to real-world tasks.

Lack of modeling of human mental model There is now a substantial body of work that emphasizes calibrating users' mental models [13, 20]. However, to calibrate the user's mental model, an essential premise is to know clearly what the user's current mental model is. Taking user trust calibration as an example, many studies have realized that people should increase trust when the AI makes correct predictions, and decrease trust when the AI makes mistakes, but they have not studied the trust level of users at the moment. If the user's trust in AI is already very low, there is no need to try to reduce the user's trust by providing additional information such as an explanation or low confidence.

A limited exploration and exploitation of design space On the one hand, the design space has not been fully exploited. A lot of design elements in the existing design space have not been fully investigated, such as modality and interactivity. In addition, there is relatively little knowledge of how the combination of different dimensions can affect users and how different dimensions interplay with each other. On the other hand, the existing design space is far from being fully explored. In each dimension, there can be plenty of new design elements. Also, there are many other possible design dimensions worth exploring.

Limited human-AI cooperation mode in XAI study Most XAI research now focuses on decision-making or recommendation scenarios. We realize that there are many other cooperation modes, such as multi-round collaboration mode and long-term collaboration mode, which require the design of XAI to consider changes in people's mental models over time, as well as the long-term impact of XAI on people.

Lack of exploration on the role of XAI in other human-centered aspects The explainability of AI can help address other aspects of human-centered problems. For example, XAI can facilitate assessments of AI security, audit fairness, design more private AI, and explore accountability issues. However, we find that existing XAI research pays less attention to these aspects.

4.4.2 Future research opportunities. Exploring human-centered XAI design in more actual scenario. Researchers can design XAI systems/interfaces for real-world applications. In an alternative way, researchers can try to recruit actual stakeholders for their research prototype. This can not only enrich the current XAI research space but also bring unique insights and practical implications for a specific type of user.

Design methods to model human mental model. Since the human mental model can determine the effects of explanations, modeling users' mental model during the interaction process is promising. At the same time, we also recognize its difficulty, because it is difficult to quantify the user's mental model. Explicit modeling methods, such as asking the user to answer a few questions to measure the user's mental model during the execution of the task, will seriously damage the user's experience. While implicit methods (such as modeling the user through the user's interaction data in the interface, the time of hesitation, or the user's expression, etc.) are likely to face the problem of inaccuracy. We believe that this is a very worthy research direction in the future.

Exploring new XAI design. The purpose of summarizing the existing design space is not to limit the ideas of designers, but to let designers realize the insufficiency and blankness of the current design space. In the future, researchers should explore new design dimensions or explore new design elements or combinations between different design elements. And empirical research can serve as the ground for exploration.

Investigating the effects of XAI beyond one-round human-AI interaction. Most of the current XAI research uses a one-round of human-AI interaction mode, or the same sub-task is repeated many times. We find that there are

still research gaps in the multi-round interaction. Researchers can explore this area in the future. There are specific research questions to explore, such as for a multi-round decision-making task, how the explanation provided by the AI should change as the interaction progresses to keep calibrate users' trust, and how the user's decision changes after seeing updated AI explanations in a new round.

Investigating the effects of XAI beyond one-on-one human-AI interaction. The current experimental scenarios are almost all one person interacting with one AI, and little work has been done to explore how multi-person and multi-AI teams interact internally. Taking decision-making tasks as an example, in a scenario of two people and two AIs, there will be many different voting situations, so how should the final decision be determined? How do team members communicate? What are the roles of humans and AI members in team decision-making? These issues are worth exploring in the future.

5 EVALUATION OF XAI DESIGN

Evaluation for XAI systems is a critical step in the XAI empirical studies. Explanations or explanatory interfaces are designed to achieve different interpretability goals (based on broad user needs or research questions), and hence different measures are needed for the intended purpose [101]. In this section, we focus on evaluation metrics. Reviewing the surveyed papers, we summarize a list of evaluation measures (*how to evaluate*) associated with their evaluation goals (*what to evaluate*), as shown in Tables 3, 4, 5.

At a high level, we group the evaluation metrics into three categories: (1) evaluation related to users' perceptions of AI, (2) evaluation related to users' interaction experience, and (3) evaluation related to the task. Under each, we categorize metrics based on evaluation goals and further classify them based on subjectivity. We note that the qualitative methods (such as interviews and think-aloud) can be diverse in each work and few works provide the question list in interviews in their papers. Thus, here we only review quantitative methods used in surveyed papers.

5.1 Evaluation Related to Users' Perceptions of AI

Since one of the most direct design goals of XAI is to help users understand the AI, a lot of evaluation is situated on this goal, including users' understanding/mental model of AI, trust/reliance in AI, trust calibration, fairness judgment, and others. Table 3 summarizes these measures.

5.1.1 Understanding/Mental Model. Users' understanding of the AI is the most investigated evaluation goal in the surveyed papers. Subjective metrics includes users' *self-report of understanding of the AI* [5, 19, 25, 31, 96, 138, 143], *confidence in understanding* [73], *ease of understanding* [111], *confidence in simulation* [3, 105], *perceived intuitiveness* [129] or *perceived transparency* [113, 133] of the AI system. Generally, in our surveyed studies, researchers often use subjective questionnaire (usually 5 or 7-point Likert scale) to measure users' understanding or users' perceived transparency of the AI [19, 43, 47, 99, 129, 133, 141].

Objective metrics is often used to test how well users understand *how the system works* and *what the output will be* compared to actual ground-truth. The most commonly used metric is forward simulation [1, 19, 32, 61, 107, 111, 138], which asks participants to guess a model's predictions on unseen instances. In a slightly different manner, counterfactual simulation is also used in some studies [61, 138]. Apart from simulation, other objective metrics has been utilized, such as the correctness of people's assessment of model performance [107, 127], identification of important features [31, 129], detection of errors [111, 138], or description of model behaviors [32]. These measures are often conducted by quizzes

[31, 51, 118, 138]. For example, Robertson et al. [118] measure users' understanding by asking participants to answer multiple-choice recall questions.

Some studies adopt both subjective measures and objective measures [31, 32, 138]. For example, in a comparison study of different types of explanation, Wang et al. [138] measure users' both subjective understanding and objective understanding. For the subjective measures, they ask participants to self-report their understanding on a Likert scale. For the objective measures, they design five kinds of quiz questions. Note that objective and subjective understanding does not always align due to some reasons, such as "illusory confidence" that humans believe they understood the model more than they actually did [31, 32].

5.1.2 Trust/Reliance. Trust in AI is an essential research topic. There are some types of subjective measures, such as direct self-reported trust [1, 19, 31, 32, 111, 127, 133], self-reported agreement or reliance [29], acceptance or confidence in the model [32, 129, 138], or perceived accuracy of the AI [70, 123, 127]. Most works in Table 3 measure users' trust subjectively. For example, Wilkinson et al. [141] (in a movie recommendation task), Buçinca et al. [19] (in a food percent fat content of nutrients recognition task), Lai et al. [79, 80] (in a deceptive review detection task), Cheng et al. [31] (in a graduate admission prediction task), measure users' trust by subjective questionnaire.

Objective metrics focus more on reliance (i.e., whether users take the model's advice or how much users' decisions are influenced by the AI's suggestions), such as acceptance of model suggestions [13, 79, 80, 93, 95, 138, 145, 148], likelihood to switch [95, 145, 148], weight of model advice [111], as well as deviation from the model's suggestions [111]. For example, in an income prediction task, Zhang et al. [148] use two indicators to measure users' trust, *switch percentage* which means how many users' decisions switch to AI's prediction, and *agreement percentage* which means how many users' decisions are aligned with AI's prediction.

It is worth noting that trust and reliance are not the same. Trust is more like an attitude and reliance is more close to a behavior. Apart from trust, reliance can be influenced by other factors, such as efforts to engage, time constraints, perceived risk, workload, and self-confidence [78, 82]. There are also some surveys that conclude how to measure users' trust in empirical studies [78, 101, 135].

5.1.3 Trust Calibration. Here, we deliberately highlight this goal separately. Although *trust calibration* can be seen as a subset of *trust*, most studies in *trust* aim to increase trust in users by providing an explanation. However, in recent days, more and more research's focus has changed from increasing users' trust to maintaining an appropriate trust in the AI. Because to achieve a complementary human-AI team performance [11, 13, 19, 114], an essential step is to guide people to trust or to be cautious in different situations with the help of explanations (trust calibration [143, 148]).

To evaluate the effects of explanation design on trust calibration, researchers usually measure it in two cases, one is whether people make decisions that are consistent with AI when AI makes correct decisions, and the other is whether people doubt AI when AI makes wrong decisions. Thus, existing empirical studies propose three objective measures, (1) the *over-reliance* [20, 138, 143] which means the human follows the AI's prediction when the AI is wrong; (2) *under-reliance* [138, 143] which means the human ignores the AI's prediction though the AI is correct; (3) *appropriate reliance* [111, 138, 143] which means an ideal situation where the human can adopt the AI's suggestion when the AI makes a correct prediction and vice versa. For example, Wang et al. [138] evaluate the effects on trust calibration by measuring users' awareness of model uncertainty and calculating the appropriate trust ratio, overtrust ratio, and undertrust ratio in the given instances.

Table 3. Evaluation related to users' perceptions of AI

Evaluation Goal	Subjectivity	Metrics
G1: Understanding/Mental Model	subjective	Self-reported understanding [5, 19, 25, 31, 138, 143], confidence in simulation [3, 105], intuitiveness [129], confidence in understanding [73], perceived transparency/interpretability [113, 133], ease of understanding [59, 111]
	objective	Forward simulation [1, 19, 32, 107, 111, 117, 138, 146], counterfactual simulation [61, 138], model errors detection [111, 138], identifying important features [31, 129], correctness of estimated model performance/accuracy [107, 123], comprehension quiz [31, 51, 73, 138], correctness of described model behaviors [32, 74, 113]
G2: Trust and Reliance	subjective	Self-reported trust [1, 19, 31, 32, 111, 127, 133], model confidence/acceptance [3, 32, 129, 138], self-reported agreement/reliance [29], perceived accuracy [70, 123, 127]
	objective	Agreement/acceptance of model suggestions [13, 79, 80, 93, 138, 145, 148], switch [145, 148], weight of advice [111], disagreement/deviation [111], choice to use the model [116]
G3: Trust Calibration	objective	over-reliance [20, 22, 138, 143], under-reliance [22, 138, 143], appropriate reliance [111, 138, 143]
G4: Fairness Judgement	subjective	Perceived fairness [5, 40, 54, 60, 134], individual/group fairness [85], deserved outcome [16], feature fairness [16, 134], accountability [113]
	objective	Decision bias [54, 55]

5.1.4 Fairness Judgement. Studying how people perceive the fairness of AI and what design impacts the perception is an active research area. These studies primarily rely on subjective metrics, from general perceived fairness [5, 40, 54, 60, 134] to perceptions of more fine-grained types of fairness such as individual fairness [85], group fairness [85], process fairness [16], deserved outcome [16], feature fairness [16, 134], and accountability (i.e., the extent to which participants think the system is fair and they can control the outputs the system produces) [113]. For example, Anik et al. [5] use a subjective questionnaire to measure users' perceived fairness of the system and the AI training. In a recidivism prediction task, Dodge et al. [40] design a questionnaire to subjectively measure users' perception of fairness. Only a small number of studies leveraged decision bias (e.g., the action to follow the model's recommendations despite their lack of fairness) [54, 55] as an objective metric of perceived fairness.

5.1.5 Other factors. One important factor that end-users are always concerned about is the privacy of their data [8, 89, 101]. An explanation is expected to facilitate privacy checking for users. Another factor is the bias in the model. A model can be biased either due to the biased training data or the biased feature engineering [8]. Explanation, especially the explanation of the training data can help deal with this problem. However, there is little explicit empirical research on these topics.

5.2 Evaluation Related to Users' Interaction Experience.

The explanation design not only affects users' perceptions of the AI itself but also affects users' experience when interacting with the explanatory system. In the user experience (UX) domain, there are rich evaluation perspectives. However, in XAI empirical studies, we mainly divide user interaction experience from three perspectives, satisfaction/perception of the system, cognitive load/mental demand, and engagement, as shown in Table 4.

5.2.1 Satisfaction/Perception of the system. Subjective metrics are often used, such as users' satisfaction with the AI [70, 76, 96, 133], perceived helpfulness/usefulness [19, 26, 70, 143], effectiveness [133], quality [133], appropriateness [19], preference [73, 83, 117, 143], etc. Besides, some studies measure system complexity [20], ease of use [5, 133], system frustration [70, 83, 123], information richness [5, 83, 84], learning effect [133], etc. Focusing on explanation, people's

Table 4. Evaluation related to users' interaction experience.

Evaluation Goal	Subjectivity	Metrics
G5: System Satisfaction/Usability	subjective	Satisfaction [17, 38, 70, 76, 133], helpfulness/usefulness [19, 26, 70, 143], effectiveness [133], quality [133], appropriateness [19], preference [73, 83, 143], complexity [20], ease of use [5, 133], system frustration [83, 123], richness/informativeness [5, 84], learning [133], recommendation to others [70], usefulness/helpfulness of explanation [26, 107, 129], easiness to use explanation [129], quality/soundness/completeness of explanation [129]
G6: Cognitive Load/Mental Demand	subjective	mental demand/effort [19, 20, 74, 140], workload [26, 84]
G7: Engagement	objective	Browsing time on the interface [47]

perceived explanation quality [74, 75, 129], explanation usefulness [26, 84, 107, 129], easiness to use explanation [129] are often subjectively measured.

For example, in an average reading time of a magazine article prediction task, Szymanski et al. [129] measure users' perception of ease-of-use of each type of explanation. In a UX problem finding task, Fan et al. [47] use a subjective questionnaire to measure users' satisfaction. And in a vocal emotion recognition task, Zhang et al. [147] measure users' perceived system helpfulness on a 7-point Likert scale. In a sentiment classification and a QA task, Bansal et al. [13] use a subjective questionnaire to measure users' perceived helpfulness. In a recidivism prediction task, Liu et al. [93] use a subjective questionnaire to measure users' perception of AI assistance's usefulness.

5.2.2 Cognitive Load/Mental Demand. On the one hand, the explanation can help users understand the AI. On the other hand, since the explanation contains extra information about the model, it might lead to cognitive load or much mental demand for users. To measure the effects of explanations, researchers have asked participants about their mental demand/effort [19, 20, 74, 140], workload [26, 84]. For example, in a task where users need to replace the highest calorie ingredient in a dish with another lower calorie but similar taste ingredients, Buçinca et al. [20] use a subjective questionnaire to measure users' mental demand and users' perception of system complexity. In a COVID-19 self-diagnosis task, Tsai et al. [133] use the NASA-TLX scale to measure users' cognitive load. In a house price prediction task, Abdul et al. [1] use a mixed measure to evaluate users' cognitive load, including the reading time, users' self-reported cognitive load, recall reconstruction score, and recognition score. In a video game event recall task, Robertson et al. [118] use a subjective questionnaire to measure users' cognitive effort.

5.2.3 Engagement. Engagement plays a crucial role in human-computer interaction. Some works also measure the effects of explanation on users' engagement. For example, in a UX problem finding task, Fan et al. [47] use the interaction logs on the interface to measure users' engagement.

5.3 Evaluation Related to Task

For the task-related measures, existing works mainly focus on two aspects, task performance, and task efficiency.

5.3.1 Task Performance. Explanations are often designed to assist users to perform tasks, thus task performance is a popular measurement in existing empirical studies.

Task performance is usually evaluated by objective measures. In classification tasks, the most commonly used metric is accuracy [11, 42, 52, 76, 79, 80, 105]. Researchers are interested in comparing the accuracy of the human-AI team

Table 5. Evaluation related to the task

Evaluation Goal	Subjectivity	Metrics
G8: Task Performance	subjective	Self-rated error/accuracy [38, 80, 133], confidence in the decisions [54, 55, 59]
	objective	Accuracy/error [11, 52, 76, 79, 80], precision [47], recall [47, 102], false positive rate [28, 42, 54], false negative rate [28, 42], mean prediction error [111]
G9: Time Spent/Efficiency	objective	time taken on the task [1, 9, 28, 31, 51, 70, 76, 121, 123, 143]

with the accuracy of human-alone or AI-alone. For example, in a food percent fat content of nutrients recognition task, Buçinca et al. [19] calculate the percentage of correct answers to measure the task performance. In an income prediction task, Zhang et al. [148] use accuracy to measure performance. In a cooking video activity recognition task, Nourani et al. [107] use the number of errors to measure the task performance.

Besides accuracy, other metrics in classification tasks are also used to evaluate the task performance, such as F1 [17], precision [17], recall [17], AUC-ROC [42], false positives rate [28, 42, 54], false negatives rate [28, 42], true positive rate [42], true negative rate [42], etc. For example, in a UX problem-finding task, Fan et al. [47] use accuracy, precision, and recall of UX problem detection to measure task performance.

In addition to objective metrics, subjective metrics are used to measure human perception of task performance, such as the perceived accuracy (i.e., self-rated error/accuracy) [38, 80, 133], humans' confidence in the decisions [54, 55, 59]. Compared to objective measures, subjective metrics are less intuitive and used.

5.3.2 Time Spent/Task Efficiency. Another task-related dimension is efficiency, which means how efficiently the user can complete the task with the AI's assistance. The most common objective metric is time spent on the task [1, 28, 31, 51, 70, 76, 121, 123, 143]. For example, in a vocal emotion recognition task, Zhang et al. [147] measure the efficiency by the logged task times for different pages. In a graduate admission prediction task, Cheng et al. [31] collect users' spent time when interacting with different explanatory interfaces.

5.4 Summary and Thinking

In this section, we review the commonly used evaluation metrics in empirical studies.

5.4.1 Challenges. Lack of standard evaluation system From the surveyed paper, we find that existing empirical studies have not yet formed a unified set of evaluation methods. Each work has adopted different evaluation methods from each other, which hinders the collaborative progress of the field and the mutual utilization of results. Many studies even directly use evaluation methods designed by themselves. Such a phenomenon is mainly due to the fact that the empirical research on XAI is still in the preliminary stage, and a standard evaluation system has not yet been formed.

5.4.2 Future research opportunities. Developing an evaluation system for XAI empirical studies. We believe that the next step in this field is to form a standard evaluation system. First, it requires a detailed division of different XAI tasks, user categories, experiment scales, goals, etc. In Section 5, we only divide the evaluation methods based on evaluation goals, which is not enough for an evaluation system because a number of other factors need to be considered. Second, the evaluation system must be scalable, because the research on XAI is just at a young age, and the existing evaluation methods are only some preliminary exploration. Therefore, there will be many excellent and appropriate evaluation methods for each kind of XAI in the future, thus the system needs to flexibly absorb new methods, and

discard or improve the old methods. Finally, to build such a system, all researchers need to work together, and some open-source and collaborative methods can be considered to maintain the evaluation system.

6 FINDINGS AND PITFALLS DERIVED FROM EMPIRICAL STUDIES

From carefully organized user studies, the empirical research of XAI has obtained valuable findings from the results. These findings can help researchers or practitioners understand the possible effect of different XAI designs on a specific user group in a specific task scenario, and help researchers get references when designing new explanations or new empirical studies. However, we find that in the previous surveys on XAI, there is scarce work to summarize the relevant literature from the perspective of findings. There are two possible reasons. First, existing XAI surveys do not focus on the perspective of empirical study [15, 63, 65, 101, 131]. Second, current works on XAI empirical study often focus on a specific task, propose a specific (set of) explanation design, verify it on a specific type of user, and use a specific evaluation method. These make it hard to make a fair comparison between the findings obtained from the different studies. We acknowledge that the findings obtained in different experimental settings may not be generalized to other experimental settings. However, we believe that analysis and summarization of the findings will help readers have a deeper understanding of the XAI design mentioned earlier in this survey.

In order to understand the current findings in the XAI empirical studies from an overall perspective, we use the qualitative analysis method commonly used in HCI to conduct a thematic analysis of the findings in the literature and extract some commonalities. We believe that the themes in these findings are of great value to researchers and practitioners, such as establishing an overall understanding of the existing work, building more comprehensive and reasonable expectations for their own experimental design, and being aware of possible pitfalls and failures, controlling confounding design factors. Overall, we categorize these findings into common findings and pitfalls, and the latter focuses on the current negative effects of XAI design. Next, we will summarize the findings based on themes. Under each theme, we will list several representative empirical studies as examples. Note that we do not list “well-known positive” findings that support the effectiveness of the explanation, such as appropriate XAI design can help users understand the model or can improve users’ satisfaction with the AI system.

6.1 Common findings

6.1.1 A moderate granularity of explanation could benefit users’ understanding of AI. There exists a trade-off between complex detailed explanations and simple explanations. For example, Mishra et al. [99] find that a balance between coarse and fine-grained explanations lead to better users mental model of the model’s predictions. Abdul et al. [1] correlate the complexity of visual explanations with the cognitive load of the user, and propose a method COGAM to trade-off between cognitive load and accuracy.

Apart from designing a moderate granularity of explanation, some empirical studies suggest a progressive disclosure of explanations. For example, Springer et al. [127] find that users can benefit from progressive disclosure of explanations where simplified feedback that helps users build heuristics about the system is shown initially. Similarly, Nourani et al.’s [107] study suggests that before using the detailed instance-level explanations, a higher-level explanation could help users build more accurate mental models.

6.1.2 Interactive explanation may and may not increase users’ understanding or task performance. The interactive explanation has been proposed in visualization and HCI domains, which is expected to enhance users’ exploration and

thus understanding of the AI model. However, through empirical studies, there are mixed findings of the effects of interactive explanation on users' understanding.

On the one hand, Cheng et al. [31] compare the effects of interactive explanations on users' understanding, trust, and time spent. They conduct the user study on a graduate student admission prediction task via 199 AMTurk participants and find that compared to the static interfaces, interactive interfaces increase the understanding of the algorithm. For data scientists, Narkar et al. [103] propose an interactive explanatory tool for multi-level model comparison in autoML, and find that participants gave high ratings to the usability of this tool. On the contrary, in a crowdsourcing user study with two types of tasks, recidivism prediction, and profession prediction. Liu et al. [93] try to understand the effect of interactivity of explanation and data distribution on human-AI decision making. They find that interactive explanations may reinforce human biases and lead to limited performance improvement.

6.1.3 Expertise matters how users perceive an explanation. It has been widely recognized that users' expertise will play a critical role in how users perceive an explanation and the effectiveness of an explanation design [5, 43, 101, 120, 128, 129, 143]. The expertise includes both AI expertise and domain expertise [101, 128]. Usually, experts users could gain more from explanations than novices. For example, Szymanski et al. [129] find that AI novices gain less benefit from explanations but are also more likely to have illusory satisfaction compared to experts. Fan et al. [47] suggest that compared with non-experts, such as crowdworkers, expert users tend to have higher confidence in their judgments and might adopt different strategies in human-AI collaboration [46, 124]. For a decision-making scenario, Wang et al. [138] compare different types of explanations in AI-assisted decision-making. They conduct a user study on two types of tasks, forest cover prediction (users lack the expertise) and recidivism prediction (users have the expertise). They find that users' expertise in different decision-making tasks highly influences the effectiveness of different XAI methods.

6.1.4 Apart from expertise, users' other intrinsic characteristics can affect their perceptions of explanations. Recent empirical studies find that there are some other users' characteristics that could affect the effects of an explanation, such as users' general trust in AI [40], locus of control [119], visual literacy [1, 18]), cognitive load disposition [52], need for cognition [24] (a personality trait reflecting one's general motivation to engage in effortful mental activities, such as thinking). Also, Anik et al. [5] suggest that users' perceptions can be influenced by their prior exposure to the AI concepts, such as what they have seen in the media. Also, humans' prior intuition is probably an informative distinguishing factor that can affect the human perception and interpretation of the explanations [80].

Taking the need for cognition as an example, some studies suggest that people are less able to process explanations effectively if the time and cognitive resources are constrained [118, 142]. Ghai et al. [52] show people who have a low score in need for cognition will be less satisfied when adding explanations in an active learning setting. To reduce users' overreliance on AI, Bućinca et al. [20] propose three kinds of cognitive forcing functions in a crowdsourcing study. Their results show that the proposed cognitive forcing interventions benefit participants who get a higher score in need for cognition more. Besides the cognitive factor, Dodge et al. [40] conduct an empirical study to explore how explanations affect users' fairness judgment. Through a crowdsourcing user study of a recidivism prediction task, they find that how users react to different styles of explanation will be influenced by some individual differences, such as their prior positions and judgment criteria of algorithmic fairness.

6.1.5 Different types of explanation could benefit different design goals. A large number of empirical studies have been designed and conducted to understand how users perceive, process, and use different types of explanation [40, 79, 103, 133, 138, 143, 147]. Generally, in different scenarios, for different objectives, one type of explanation could

be more effective than another. For example, Dodge et al. [40] explore the effect of different types of explanations on calibrating users' perceived fairness of ML models. They find that compared to global explanations, local explanations are more effective in calibrating users' fairness judgment of ML models, because local explanations can highlight unfair features used for individual predictions. In a different setting, Tsai et al. [133] design three types of explanations for online symptom checkers. Through a lab-controlled user study, they find that users may have different preferences for different explanations. Static or "one-size-fits-all" explanations cannot fulfill users' needs as they would like to get some control and customization in what explanations are conveyed to them. These findings imply that the explanations are expected to be adaptive and controllable based on the users' real-time information needs.

6.1.6 Tutorial and introduction of the model can benefit users' understanding. Sometimes, to achieve an effective interaction with AI, users need to access the global information of an AI system to build an appropriate understanding of the underlying model and data. Recently, some works provide documentation to present global information to users (e.g., Model cards [100], Datasheet [50], FactSheets [7]). Empirical studies have shown that offering users a tutorial or an introduction to the whole model or data can lead to a more accurate mental model. For example, Lai et al. [79] propose some types of tutorials of ML models, and compare them with no-tutorial conditions in a deceptive review detection task with MTurk participants. They find that tutorials can indeed improve human performance to some extent.

6.2 Pitfalls

6.2.1 Transparent systems do not necessarily lead to high trust or better task performance. Some explanations are designed to improve users' trust and task performance based on an assumption that a transparent system will help users understand the AI and thus improve users' trust and task outcomes. However, several studies find that explanations could lead to some negative effects, such as decreasing users' situation awareness and leading to worse task performance [35, 111, 120, 148], because explanations could overwhelm or overload users with much information about the system.

For example, Robertson et al. [118] investigate different behavior explanations for a real-time strategy game setting and find that a *why* explanation does not improve users' task performance because users lack enough cognitive resources to interpret the *why* answers in real-time. In a student admission prediction task, Cheng et al. [31] find that showing the explanation does not increase users' trust in the algorithm compared with a black-box model without explanation. Bansal et al. [13] design an empirical study to investigate the effects of explanations on human-AI team performance. They find that explanations do not increase team performance but increase the chance that users follow the AI's suggestion, regardless of its correctness. And they suggest that "instead of convincing, explanations should be informative". Besides, some recent empirical studies investigate the effects of the level of transparency on users' trust in the AI system [31, 75, 111, 120]. They manipulate the model configuration to increase the transparency of the model by providing explanations, reducing the feature number, and allowing users to inspect the model behavior, using a white-box model. However, they find no significant improvement in users' trust. For example, in an apartment price prediction task, they do not find that participants follow the transparent model's predictions when the model makes a correct prediction. Furthermore, showing participants a clear model hinders their ability to detect and correct the mistakes of the model, which is possibly due to information overload.

6.2.2 Human-AI teams with comparable performance do not guarantee complementary performance. A common XAI scenarios is decision-making, where the human makes the final decisions based on the AI's suggestions and explanations. Hence, the human-AI team's performance, i.e., the accuracy of the final decision, is regarded as an essential factor to measure the effectiveness of the XAI design. In recent works, researchers begin to investigate whether the human-AI

team can achieve a complementary performance when the human and AI have comparable performance. However, some evidence shows that comparable performance alone cannot guarantee complementary performance [13, 93, 114, 148].

For example, from a crowdsourcing study on an income prediction task, Zhang et al. [148] find that although a confidence score can help calibrate people’s trust in an AI model, it is still not sufficient to improve team performance. The joint team performance may also depend on whether the human and AI have non-overlapping knowledge and different error boundary. In a similar view, Bansal et al. [13] emphasize that what plays an important role is the complementary knowledge between humans and AI. Rather, in an ideal team, humans and AI could maximize their talents in different dimensions. For example, for clinical decision-making, AI could quickly sort similar cases from the database based on its computing power, and the doctor could make a unique diagnosis based on her rich domain knowledge [132].

6.2.3 Too technical or visual-complicated explanations might be hard to understand, even for experts. Although explanations are usually designed for users’ understanding. However, some explanation methods are technique-centered rather than experience-centered [89], which hinders users’ ability to correctly or fully understand them. The information contained in sophisticated explanations can also lead to information overload and prevent people from forming a proper mental model [127]. For example, Zhang et al. [147] propose a framework to generate different types of relatable explanations and in their study, they find that saliency visualization is not useful since it is too technical and complicated. Yang et al. [143] focus on visual explanations and explore the effects of spatial layout and visual representation. They find that visual explanations can lead to inappropriate trust if an explanation is difficult to understand. Generally, expert users have a stronger ability to analyze complex explanations. For example, Szymanski et al. [129] find that although lay users prefer visual explanations, they get significantly worse performance with it. And they find experts can better understand visual and textual explanations. However, it should be noted that prior research shows that even ML experts face challenges to interpret them correctly without assistance [32, 67].

6.2.4 Cognitive biases can affect users’ mental model of XAI systems and can lead to negative effects. Based on cognitive science, some empirical studies have looked into how cognitive biases exist and affect users’ perceptions of an explanation. And they reveal that cognitive biases can prevent XAI from doing what it was designed to do. For example, Nourani et al. [107] investigate the effects of anchoring bias on users’ mental model formation. And they find that the first impression plays a key role in users’ perception of AI. Specifically, users who encounter system strengths early are more prone to automation bias and make significantly more errors due to positive first impressions. On the contrary, users who observe system weaknesses early make significantly fewer errors because they tend to rely more on themselves, while the negative first impression also makes them underestimate the capability of the model. Chromik et al. [32] compare a moderated interaction with AI and an unmoderated interaction with AI, and explore the effects of lay users’ perceived understanding of the AI. They find that in the unmoderated setting, participants often adopt heuristic thinking and cannot realize the incompleteness of their understandings until they see their test results.

Based on theories of cognitive science and psychology, researchers have examined humans’ cognitive processes when they interpret AI explanations. One of the most representative theories mentioned in existing empirical studies is the dual-process of cognition [19]. From the dual-process theory [23, 66, 139], humans’ cognitive processes can be driven by two systems: System 1 and System 2. The former makes humans think fast and process information in an automatic manner, whereas the latter makes humans think slow and engaged in deliberative and analytical thinking. System 1 usually relies on developed heuristics (rules-of-thumb or mental short-cuts) which could lead to cognitive biases if applied inappropriately [66]. Based on this theoretical foundation, there is an increasing awareness [19, 43, 107, 114, 136]

that while XAI developers assume that humans will carefully digest every information in explanations (analytic System 2 thinking), in reality, people are more likely to adopt System 1 thinking as it is fast and labor-saving. Cognitive biases can lead to many negative consequences such as over-trust in XAI (mentioned below), which can be attributed to users' associating explanations with AI competence in System 1 heuristic [32, 107].

6.2.5 Explanation can lead to users' over-trust even the model or the explanation is not "good". Some recent works show that explanations, even if they are placebic or randomly generated, may improve humans' trust in AI predictions [13, 54, 55, 80]. For example, Kaur et al. [67] find that the existence of explanations can make data scientists hold an over-confidence and mistakenly think that the model is ready for deployment. In addition, there is the concern of illusory understanding which means that humans can over-estimate their understanding gained from XAI [32]. For the underlying mechanism of why XAI could lead to over-reliance, some studies find that explanations are generally interpreted as a signal of competence, no matter what content is in the explanation, and just the presence of an explanation can increase humans' trust in the AI [13, 52, 111, 148]. As mentioned in the above section, humans are easily engaged in System 1 thinking so that they simply associate explanation with AI's capability without engaging analytically with the model behaviors.

Besides the explanation, other types of explanatory elements could also lead to over-trust issues. For example, Lai et al. [80] find that participants are more likely to trust models with accuracy statements than models without accuracy statements, even if poor performance is stated. These observations are consistent with prior work on numeracy which suggests that it is hard for humans to interpret and act on numbers [14, 110, 115, 122]. Also, Lai et al. [80] show that adding random heatmap as explanations can enhance humans' trust. Such an over-reliance phenomenon does not just happen to non-expert users. Some papers find that both expert users and non-expert users have unwarranted faith in numbers. For example, Ehsan et al. [43] find that participants in both AI-expert and AI-novice groups have unwarranted faith in numbers. To mitigate these issues, recent studies find that presenting information about uncertainty is an effective means to help users maintain appropriate trust in AI [13, 114, 148].

6.3 Summary and Thinking

In this section, we summarize the common findings and pitfalls of current XAI design, which can offer valuable implications for researchers to develop new XAI interfaces and conduct new empirical studies.

6.3.1 Challenges. The experimental conclusions are diverse and difficult to generalize Although we have summarized some common experimental findings in Section 6, we find that diversity exists in existing experimental findings. For example, some studies find that interactive XAI design helps to enhance users' trust in AI, while some works find that interactive XAI does not have such an effect. There are many reasons behind this, such as different XAI design details, different users, different tasks, different experimental procedures, different AI algorithms, different evaluation metrics, and more. Although the diversity of research results is beneficial for the healthy development of the young field, it prevents the generalization of different studies and hinders the joint efforts of different research. However, we suggest that when different works analyze their experimental conclusions, they can generalize the experimental conclusions to a more general level by analyzing the reasons and patterns behind them based on relevant theories, which will help researchers to establish a systematic understanding in this field.

Lack of successful application cases Although the explanation of recommender systems has been widely adopted, and many products now have a simple explanation of the newly added AI functions, we still rarely see the success of other AI tools in our lives, such as a decision support system. On the one hand, this stems from the lack of attention to

users' needs for interpretability. Many AI products still only focus on providing functions and do not pay attention to users' needs for interpretation of the functions provided. On the other hand, providing any explanation will actually bring extra information to the user, which will affect the existing task flow, and may also be inconsistent with the mental model that the user has established before. If not well designed, the steps in which the user interacts with the explanation interface may degrade the user's user experience, and even the explanation provided does not really help the user understand the system. In short, in order to promote the application of XAI in real life and work, the cooperation between academia and industry, as well as the cooperation of multidisciplinary researchers is needed.

6.3.2 Future research opportunities. Developing more theory-grounded XAI interface and empirical studies. From the analysis of existing studies, we find many findings can echo well-known theories from social science, psychology, cognitive science. Thus, we may leverage explanation-related theories, such as how a human explains to another human, to guide more effective XAI design and more scientific empirical study.

Designing more effective XAI methods for human-AI decision-making. As mentioned in this section, XAI has the potential to help users build more accurate mental models of the AI and achieve complementary human-AI team performance in decision-making. However, we also notice that the current explanation design is far from successful due to a lack of consideration of humans' cognitive processes. In the future, we can focus on specific decision-making tasks, and leverage XAI to mitigate the potential cognitive biases.

7 A FRAMEWORK FOR HUMAN-CENTERED DESIGN OF XAI WITH EMPIRICAL STUDIES

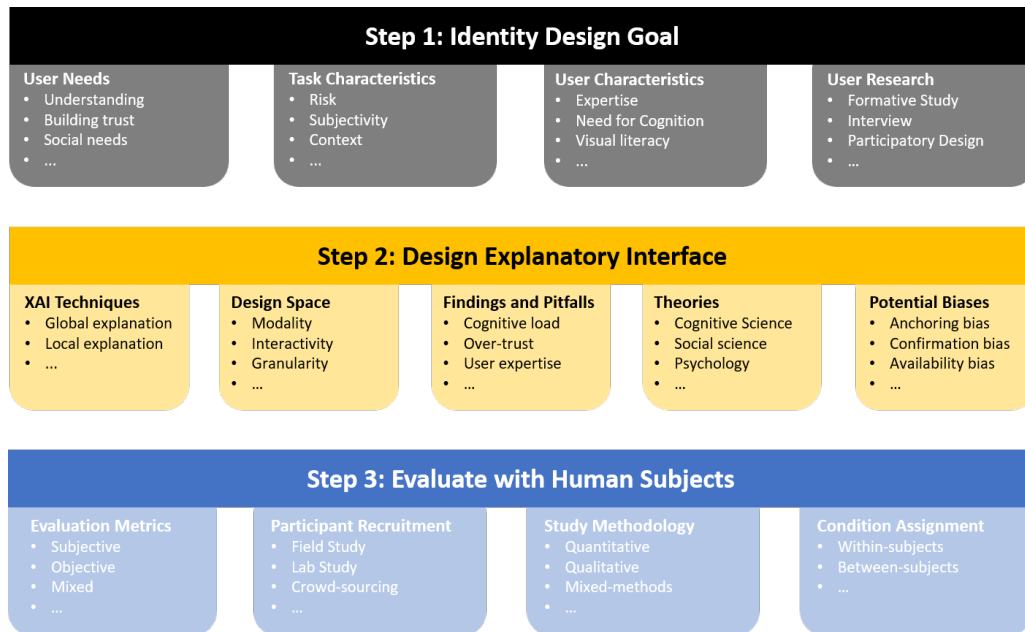


Fig. 4. A framework of human-centered XAI design.

From the survey papers, we can see a lot of good practices in designing human-centered XAI with empirical studies. Besides their specific design and unique findings, their successful XAI design and user study process are also valuable

and insightful for researchers in this area. Thus, based on the lessons learned from the surveyed papers, we propose a framework for the human-centered design of XAI with empirical studies. The framework is illustrated in Figure 1. Overall, there are three stages in the whole procedure, (1) Identify Design Goal, (2) Design Explanatory Interface, (3) Evaluate with Human Subjects.

7.1 Identify Design Goal

The first step for human-centered XAI design is to identify the design goal. Generally, the design goal needs to consider user needs and task characteristics. Users' needs for explainability can be quite diverse, such as understanding the AI, building trust, judging the model's fairness, getting informative decision support, etc [87, 89, 101]. As described in Section 3, designers can obtain user needs based on user characteristics or user research. There have been some well-established user group categorizations, such as dividing users into AI experts, domain experts, non-experts based on expertise [101]. And designers can adopt a user group-driven explainability needs finding method, to determine the possible user needs based on which category the target users belong to. For example, if the target users are lay people who use the AI application in their daily lives, their needs might include building trust, understanding the AI, etc. Designers can also adopt a user research-driven explainability needs finding method. Also, they can conduct a formative study with the target populations, interview the actual users, invite stakeholders to participate in the prototype design process, and more. From the user research, designers can get more detailed user needs and nuanced implications for the next design step.

Another factor to be considered for identifying the design goal is the task characteristics. In a recent paper, Lai et al. [78] categorize the AI-assisted tasks based on application domains, including Law&Civic, Medicine&Healthcare, Finance&Business, Education, Leisure, Professional, Artificial, Generic, and others. Besides the domain-based categorization, they also highlight four dimensions that are critical to distinguishing AI-assisted tasks: (1) task risk (e.g., high, low, or artificial stakes), (2) required expertise in the task, and (3) decision subjectivity, and (4) AI for emulation vs. discovery. Detailed information can be found in [78]. Different tasks will require different design goals. For example, a high-stake task in the medicine domain, such as medical disease diagnosis, may require the XAI design to be responsible within strict regulation. While a low-stake task in leisure, such as music recommendation, may require the XAI design to increase user experience.

We recommend that the designers take both user needs and task characteristics into consideration. They can first leverage the user group-driven needs finding method to obtain some alternative user needs, then based on these alternatives design appropriate user research to find the exact user needs in the specific task.

7.2 Design Explanatory Interface

The second step for human-centered XAI design is to design the explanatory interface. To begin with, designers need to choose an appropriate XAI technique (in Section 2) based on the specific AI algorithms used in the task and design goals obtained from the first step. Only determining the XAI technique is not enough as the explanatory interface and interaction design can critically affect users' perceptions and the effects of the explanation. Thus, the next step is to decide the design choice from the wide design space as shown in Section 4. To select a specific type of design, apart from focusing on achieving the design goal, we provide three recommendations.

First, designers should be aware of the common findings and pitfalls of different XAI designs from existing empirical studies. This step is not to say that the existing findings and pitfalls of different designs are necessarily applicable to the current task, but to help designers have a general understanding of the potential impact of different design methods on

users, so as to guide them in making more reasonable design. For example, if the designer is to design an explanatory interface for an AI-assisted decision-making task, being able to know that providing explanations from AI might lead to users' over-reliance problems can be a great help for them to make better design choices.

Second, designers can refer to XAI-related theories, such as sociology, cognitive science, psychology, and more. Because XAI itself involves *how users receive explanation information*, *how users process explanation information*, and *how users react to specific explanation*, this series of processes will be affected by people's inherent cognitive processes and mental models. Therefore, knowing the basic theoretical knowledge helps to design interpretable interfaces that can conform to the user's cognitive processes and mental models. In recent years, more and more works have recognized the importance of theory-based XAI design [89, 98, 136]. For example, Liao et al. [89] highlight the importance of theoretical analysis of human explanations, cognitive and behavioral processes. Wang et al. [136] propose a conceptual framework to help designers to map users' reasoning needs to XAI methods. The framework involves four dimensions that describe how humans reason with explanations, including explanation goals, reasoning process, causal explanation type, and elements in rational choice decisions. Miller et al. [98] summarize four major properties of human explanations from a lens of philosophy, psychology, and cognitive science. They find that explanations are often contrastive, selected, and social, and probabilities or statistical explanation can be ineffective. Designers can draw inspiration from these theories.

Third, designers should keep in mind users' potential cognitive biases, especially when designing explanatory interfaces for decision-making tasks. One direct reason for the cognitive bias is that people often make decisions fast with heuristic, which is best described with the dual-process model [66]. Users tend to adopt their System 1 thinking in making a decision, which employs heuristics and shortcuts and could lead to cognitive biases. Thus, designers might need to design the explanatory interface while considering whether it can make users engage analytically with the explanations. For example, Wang et al. [136] propose several explanation design methods to mitigate different kinds of cognitive biases, including (1) mitigating representativeness bias by prototyping cases of decision outcomes, (2) mitigating availability bias by showing the prevalence of decision outcomes, (3) mitigating anchoring bias by premortem of decision outcome, (4) mitigating confirmation bias by discouraging backward-driven reasoning, (5) facilitating moderate trust by exposing system state and confidence.

7.3 Evaluate with Human Subjects

The third step for human-centered XAI design is to evaluate the design with human subjects. First, designers need to determine the evaluation metrics based on the evaluation goals, e.g., users' mental model, task performance, etc. These metrics can be either subjective or objective or mixed based on the specific evaluation goal and the availability of user data, as shown in Section 5. Apart from the evaluation metrics, designers are required to decide participant recruitment method, determine the study methodology, and design condition assignment.

The participant recruitment method includes lab study, crowd-sourcing, online study, field study, etc. From the surveyed empirical studies, we find that nearly half of them recruit non-expert participants via crowd-sourcing platforms such as Amazon Mechanical Turk. The required expertise of participants can determine the participant recruiting method and number of participants [33]. Recruiting difficulty is likely to increase with the required level of participants' expertise [41]. One can recruit novices in large numbers via crowd-sourcing. In contrast, domain or AI experts are usually harder to recruit. They are often invited to a targeted online study, a lab study, or a field study.

The study methodology usually follows a qualitative, quantitative, or mixed study approach. The choice can be influenced by the evaluation goal and the participant recruitment method. For example, qualitative methods or mixed methods are more common in in-lab studies, while quantitative methods are more often seen in crowd-sourcing studies.

Condition assignment should be decided by target evaluation goals. Between-subjects designs investigate the differences between groups of participants, each usually assigned to one XAI condition. In contrast, within-subject designs study differences within individual participants who are assigned to multiple XAI conditions. Also, note that order effects should be considered in a within-subject design.

These are common factors of a user study in the HCI domain, not limited to the empirical study of XAI, thus in this survey, we do not focus on these aspects. However, they are essential components in the whole framework.

7.4 Summary

Designers can use this framework to perform a conceptual analysis of 1) how to identify the design goal based on user needs research, user characteristics, task characteristics, and user research; 2) how to design the explanatory interface by considering XAI techniques, design space, common findings and pitfalls obtained from existing work, theories, and potential biases; 3) how to evaluate with human subjects by selecting evaluation metrics, designing appropriate participant recruitment, determining study approach and condition assignment. Although some aspects of the framework are not covered in this survey, we provide recommended references for interested readers to check. We hope this draft framework can help designers build a comprehensive picture of what the process of a human-centered XAI design might look like.

REFERENCES

- [1] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y Lim. 2020. COGAM: measuring and moderating cognitive load in machine learning model explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.
- [3] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 275–285.
- [4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.
- [5] Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [6] Suresh Kumar Annappindi. 2014. System and method for predicting consumer credit risk using income risk based credit score. US Patent 8,799,150.
- [7] Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. 2019. FactSheets: Increasing trust in AI services through supplier’s declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6–1.
- [8] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bernetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.
- [9] Syed Z Arshad, Jianlong Zhou, Constant Bridon, Fang Chen, and Yang Wang. 2015. Investigating user confidence for uncertainty presentation in predictive decision making. In *Proceedings of the annual meeting of the Australian special interest group for computer human interaction*. 352–360.
- [10] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012* (2019).
- [11] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.
- [12] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2429–2437.
- [13] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [14] Donald M Berwick, Harvey V Fineberg, and Milton C Weinstein. 1981. When doctors meet numbers. *The American journal of medicine* 71, 6 (1981), 991–998.

- [15] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 648–657.
- [16] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*. 1–14.
- [17] Or Biran and Kathleen R McKeown. 2017. Human-Centric Justification of Machine Learning Predictions.. In *IJCAI*, Vol. 2017. 1461–1467.
- [18] Jeremy Boy, Ronald A Rensink, Enrico Bertini, and Jean-Daniel Fekete. 2014. A principled way of assessing visualization literacy. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 1963–1972.
- [19] Zana Bućinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th international conference on intelligent user interfaces*. 454–464.
- [20] Zana Bućinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.
- [21] Margaret Burnett. 2020. Explaining AI: fairly? well?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 1–2.
- [22] Adrian Bussone, Simone Stumpf, and Dymrna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*. IEEE, 160–169.
- [23] John T Cacioppo and Richard E Petty. 1984. The elaboration likelihood model of persuasion. *ACR North American Advances* (1984).
- [24] John T Cacioppo, Richard E Petty, and Chuan Feng Kao. 1984. The efficient assessment of need for cognition. *Journal of personality assessment* 48, 3 (1984), 306–307.
- [25] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces*. 258–262.
- [26] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–14.
- [27] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.
- [28] Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2020. Feature-Based Explanations Don't Help People Detect Misclassifications of Online Toxicity. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 95–106.
- [29] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. 2018. Do explanations make VQA models more predictable to a human? *arXiv preprint arXiv:1810.12366* (2018).
- [30] Furui Cheng, Dongyu Liu, Fan Du, Yanna Lin, Alexandra Zyttek, Haomin Li, Huamin Qu, and Kalyan Veeramachaneni. 2021. VBridge: Connecting the Dots Between Features and Data to Explain Healthcare Models. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 378–388.
- [31] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.
- [32] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I think i get your point, AI! the illusion of explanatory depth in explainable AI. In *26th International Conference on Intelligent User Interfaces*. 307–317.
- [33] Michael Chromik and Martin Schuessler. 2020. A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI. *ExSS-ATEC@ IUI* 94 (2020).
- [34] Ivan Contreras, Josep Vehi, et al. 2018. Artificial intelligence for diabetes management and decision support: literature review. *Journal of medical Internet research* 20, 5 (2018), e10775.
- [35] Mary Cummings. 2004. Automation bias in intelligent time critical decision support systems. In *AIAA 1st intelligent systems technical conference*. 6313.
- [36] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of Data and Analytics*. Auerbach Publications, 296–299.
- [37] Amit Dhurandhar, Karthikeyan Shanmugam, and Ronny Luss. 2020. Enhancing Simple Models by Exploiting What They Already Know. In *International Conference on Machine Learning*. PMLR, 2525–2534.
- [38] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64, 3 (2018), 1155–1170.
- [39] Steven E Dilsizian and Eliot L Siegel. 2014. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Current cardiology reports* 16, 1 (2014), 1–8.
- [40] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*. 275–285.
- [41] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).
- [42] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaa05580.
- [43] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I Lee, Michael Muller, Mark O Riedl, et al. 2021. The who in explainable AI: how AI background shapes perceptions of AI explanations. *arXiv preprint arXiv:2107.13509* (2021).

- [44] Upol Ehsan and Mark O Riedl. 2020. Human-centered explainable ai: towards a reflective sociotechnical approach. In *International Conference on Human-Computer Interaction*. Springer, 449–466.
- [45] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O Riedl. 2021. Operationalizing human-centered perspectives in explainable AI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–6.
- [46] Mingming Fan, Ke Wu, Jian Zhao, Yue Li, Winter Wei, and Khai N Truong. 2019. VisTA: Integrating machine intelligence with visualization to support the investigation of think-aloud sessions. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 343–352.
- [47] Mingming Fan, Xianyou Yang, Tsz Tung Yu, Vera Q Liao, and Jian Zhao. 2021. Human-AI Collaboration for UX Evaluation: Effects of Explanation and Synchronization. *arXiv preprint arXiv:2112.12387* (2021).
- [48] Shi Feng and Jordan Boyd-Graber. 2019. What can ai do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 229–239.
- [49] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2019. All Models are Wrong, but Many are Useful: Learning a Variable’s Importance by Studying an Entire Class of Prediction Models Simultaneously. *J. Mach. Learn. Res.* 20, 177 (2019), 1–81.
- [50] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.
- [51] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R Millen, Murray Campbell, et al. 2020. Mental models of ai agents in a cooperative game setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [52] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–28.
- [53] Soumya Ghosh, Q Vera Liao, Karthikeyan Natesan Ramamurthy, Jiri Navratil, Prasanna Sattigeri, Kush R Varshney, and Yunfeng Zhang. 2021. Uncertainty Quantification 360: A Holistic Toolkit for Quantifying and Communicating the Uncertainty of AI. *arXiv preprint arXiv:2106.01410* (2021).
- [54] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency*. 90–99.
- [55] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.
- [56] Nina Grgić-Hlača, Christoph Engel, and Krishna P Gummadi. 2019. Human decision making with machine assistance: An experiment on bailing and jailing. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.
- [57] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.
- [58] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. *Science Robotics* 4, 37 (2019), eaay7120.
- [59] Shunan Guo, Fan Du, Sana Malik, Eunye Koh, Sungchul Kim, Zhicheng Liu, Donghyun Kim, Hongyuan Zha, and Nan Cao. 2019. Visualizing uncertainty and alternatives in event sequence predictions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [60] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. 2020. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 392–402.
- [61] Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? *arXiv preprint arXiv:2005.01831* (2020).
- [62] Michael Hind. 2019. Explaining explainable AI. *XRDS: Crossroads, The ACM Magazine for Students* 25, 3 (2019), 16–19.
- [63] Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. 2018. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE transactions on visualization and computer graphics* 25, 8 (2018), 2674–2693.
- [64] Fred Hohman, Arjun Srinivasan, and Steven M Drucker. 2019. TeleGam: Combining visualization and verbalization for interpretable machine learning. In *2019 IEEE Visualization Conference (VIS)*. IEEE, 151–155.
- [65] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–26.
- [66] Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.
- [67] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: understanding data scientists’ use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.
- [68] Amir E Khandani, Adlar J Kim, and Andrew W Lo. 2010. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance* 34, 11 (2010), 2767–2787.
- [69] Amirhossein Kiani, Bora Uyumazturk, Pranav Rajpurkar, Alex Wang, Rebecca Gao, Erik Jones, Yifan Yu, Curtis P Langlotz, Robyn L Ball, Thomas J Montine, et al. 2020. Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ digital medicine* 3, 1 (2020), 1–8.

- [70] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [71] Pigi Kouki, James Schaffer, Jay Pujara, John O’Donovan, and Lise Getoor. 2019. Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 379–390.
- [72] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5686–5697.
- [73] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1–10.
- [74] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users’ mental models. In *2013 IEEE Symposium on visual languages and human centric computing*. IEEE, 3–10.
- [75] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.
- [76] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006* (2019).
- [77] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. 2019. Human evaluation of models built for interpretability. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 59–67.
- [78] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. *arXiv preprint arXiv:2112.11471* (2021).
- [79] Vivian Lai, Han Liu, and Chenhao Tan. 2020. "Why is' Chicago' deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [80] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*. 29–38.
- [81] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1675–1684.
- [82] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.
- [83] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2020. Co-design and evaluation of an intelligent decision support system for stroke rehabilitation assessment. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–27.
- [84] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2021. A Human-AI Collaborative Approach for Clinical Decision Making on Rehabilitation Assessment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [85] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.
- [86] Ariel Levy, Monica Agrawal, Arvind Satyanarayan, and David Sontag. 2021. Assessing the Impact of Automated Suggestions on Decision Making: Domain Experts Mediate Model Errors but Take Less Initiative. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [87] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [88] Q Vera Liao, Milena Pribić, Jaesik Han, Sarah Miller, and Daby Sow. 2021. Question-driven design process for explainable ai user experiences. *arXiv preprint arXiv:2104.03483* (2021).
- [89] Q Vera Liao and Kush R Varshney. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv preprint arXiv:2110.10790* (2021).
- [90] Brian Y Lim and Anind K Dey. 2009. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th international conference on Ubiquitous computing*. 195–204.
- [91] Zhiyuan "Jerry" Lin, Jongbin Jung, Sharad Goel, and Jennifer Skeem. 2020. The limits of human predictions of recidivism. *Science advances* 6, 7 (2020), eaaz0652.
- [92] Zachary C Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.
- [93] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–45.
- [94] Tania Lombrozo. 2006. The structure and function of explanations. *Trends in cognitive sciences* 10, 10 (2006), 464–470.
- [95] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [96] Ana Lucic, Hinda Haned, and Maarten de Rijke. 2020. Why does my model fail? contrastive local explanations for retail forecasting. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 90–98.

- [97] Shuai Ma, Zijun Wei, Feng Tian, Xiangmin Fan, Jianming Zhang, Xiaohui Shen, Zhe Lin, Jin Huang, Radomír Měch, Dimitris Samaras, et al. 2019. SmartEye: assisting instant photo taking via integrating user preference with deep view proposal network. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.
- [98] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.
- [99] Swati Mishra and Jeffrey M Rzeszotarski. 2021. Crowdsourcing and Evaluating Concept-driven Explanations of Machine Learning Models. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–26.
- [100] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.
- [101] Sina Mohseni, Nilofar Zarei, and Eric D Ragan. 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11, 3-4 (2021), 1–45.
- [102] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682* (2018).
- [103] Shweta Narkar, Yunfeng Zhang, Q Vera Liao, Dakuo Wang, and Justin D Weisz. 2021. Model LineUpper: Supporting Interactive Model Comparison at Multiple Levels for AutoML. In *26th International Conference on Intelligent User Interfaces*. 170–174.
- [104] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. 2016. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Advances in neural information processing systems* 29 (2016).
- [105] Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 1069–1078.
- [106] Don Norman. 2013. *The design of everyday things: Revised and expanded edition*. Basic books.
- [107] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *26th International Conference on Intelligent User Interfaces*. 340–350.
- [108] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8779–8788.
- [109] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A Slow Algorithm Improves Users’ Assessments of the Algorithm’s Accuracy. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–15.
- [110] Ellen Peters, Daniel Västfjäll, Paul Slovic, CK Mertz, Ketti Mazzocco, and Stephan Dickert. 2006. Numeracy and decision making. *Psychological science* 17, 5 (2006), 407–413.
- [111] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.
- [112] Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. 2018. Stakeholders in explainable AI. *arXiv preprint arXiv:1810.00184* (2018).
- [113] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.
- [114] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R Varshney, Amit Dhurandhar, and Richard Tomsett. 2020. Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making. *arXiv preprint arXiv:2010.07938* (2020).
- [115] Valerie F Reyna and Charles J Brainerd. 2008. Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and individual differences* 18, 1 (2008), 89–107.
- [116] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
- [117] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.
- [118] Justus Robertson, Athanasios Vasileios Kokkinakis, Jonathan Hook, Ben Kirman, Florian Block, Marian F Ursu, Sagarika Patra, Simon Demediuk, Anders Drachen, and Oluseyi Olarewaju. 2021. Wait, but why?: assessing behavior explanation strategies for real-time strategy games. In *26th International Conference on Intelligent User Interfaces*. 32–42.
- [119] Julian B Rotter. 1966. Generalized expectancies for internal versus external control of reinforcement. *Psychological monographs: General and applied* 80, 1 (1966), 1.
- [120] James Schaffer, John O’Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 240–251.
- [121] Dylan Slack, Sorelle A Friedler, Carlos Scheidegger, and Chitradeep Dutta Roy. 2019. Assessing the local interpretability of machine learning models. *arXiv preprint arXiv:1902.03501* (2019).
- [122] Paul Slovic and Ellen Peters. 2006. Risk perception and affect. *Current directions in psychological science* 15, 6 (2006), 322–325.
- [123] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. 2020. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.

- [124] Ehsan Jahangirzadeh Soure, Emily Kuang, Mingming Fan, and Jian Zhao. 2021. CoUX: Collaborative Visual Analysis of Think-Aloud Usability Test Videos for Digital Interfaces. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 643–653.
- [125] Fabian Sperrle, Mennatallah El-Assady, Grace Guo, Rita Borgo, D Horng Chau, Alex Endert, and Daniel Keim. 2021. A Survey of Human-Centered Evaluations in Human-Centered Machine Learning. In *Computer Graphics Forum*, Vol. 40. Wiley Online Library, 543–568.
- [126] Fabian Sperrle, Mennatallah El-Assady, Grace Guo, Duen Horng Chau, Alex Endert, and Daniel Keim. 2020. Should we trust (x) AI? Design dimensions for structured experimental evaluations. *arXiv preprint arXiv:2009.06433* (2020).
- [127] Aaron Springer and Steve Whittaker. 2019. Progressive disclosure: empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th international conference on intelligent user interfaces*. 107–120.
- [128] Harini Suresh, Steven R Gomez, Kevin K Nam, and Arvind Satyanarayan. 2021. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [129] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*. 109–119.
- [130] Sarah Tan, Rich Caruana, Giles Hooker, Paul Koch, and Albert Gordo. 2018. Learning global additive explanations for neural nets using model distillation. (2018).
- [131] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. 2018. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552* (2018).
- [132] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. 2019. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*. PMLR, 359–380.
- [133] Chun-Hua Tsai, Yue You, Xinning Gui, Yubo Kou, and John M Carroll. 2021. Exploring and promoting diagnostic transparency and explainability in online symptom checkers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [134] Niels Van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B Skov. 2021. Effect of information presentation on fairness perceptions of machine learning predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [135] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–39.
- [136] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.
- [137] Xinxi Wang and Ye Wang. 2014. Improving content-based and hybrid music recommendation using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia*. 627–636.
- [138] Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*. 318–328.
- [139] Peter C Wason and J St BT Evans. 1974. Dual processes in reasoning? *Cognition* 3, 2 (1974), 141–154.
- [140] Hilde JP Weerts, Werner van Ipenburg, and Mykola Pechenizkiy. 2019. A human-grounded evaluation of shap for alert processing. *arXiv preprint arXiv:1907.03324* (2019).
- [141] Darcia Wilkinson, Öznur Alkan, Q Vera Liao, Massimiliano Mattetti, Inge Vejsbjerg, Bart P Knijnenburg, and Elizabeth Daly. 2021. Why or why not? The effect of justification styles on chatbot recommendations. *ACM Transactions on Information Systems (TOIS)* 39, 4 (2021), 1–21.
- [142] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang’Anthony’ Chen. 2020. CheXplain: enabling physicians to explore and understand data-driven, AI-enabled medical imaging analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [143] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users’ appropriate trust in machine learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 189–201.
- [144] Yi Yang, Wei Qian, and Hui Zou. 2018. Insurance premium prediction via gradient tree-boosted Tweedie compound Poisson models. *Journal of Business & Economic Statistics* 36, 3 (2018), 456–470.
- [145] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.
- [146] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2019. Do i trust my machine teammate? an investigation from perception to decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 460–468.
- [147] Wencan Zhang and Brian Y Lim. 2021. Towards Relatable Explainable AI with the Perceptual Process. *arXiv preprint arXiv:2112.14005* (2021).
- [148] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.