Do People Appropriately Rely on AI-Advice? An Analytical Review of HCI Research on Human-AI Decision-Making

Muhammad Raees
Rochester Institute of Technology
Rochester, New York, USA
mr2714@rit.edu

Vassilis-Javed Khan
Independent Researcher
Brussels, Belgium
vjkhan@vjkhan.com

Ioanna Lykourentzou
Utrecht University
Utrecht, Netherlands
i.lykourentzou@uu.nl

Konstantinos Papangelis
Rochester Institute of Technology
Rochester, New York, USA
kxpigm@rit.edu

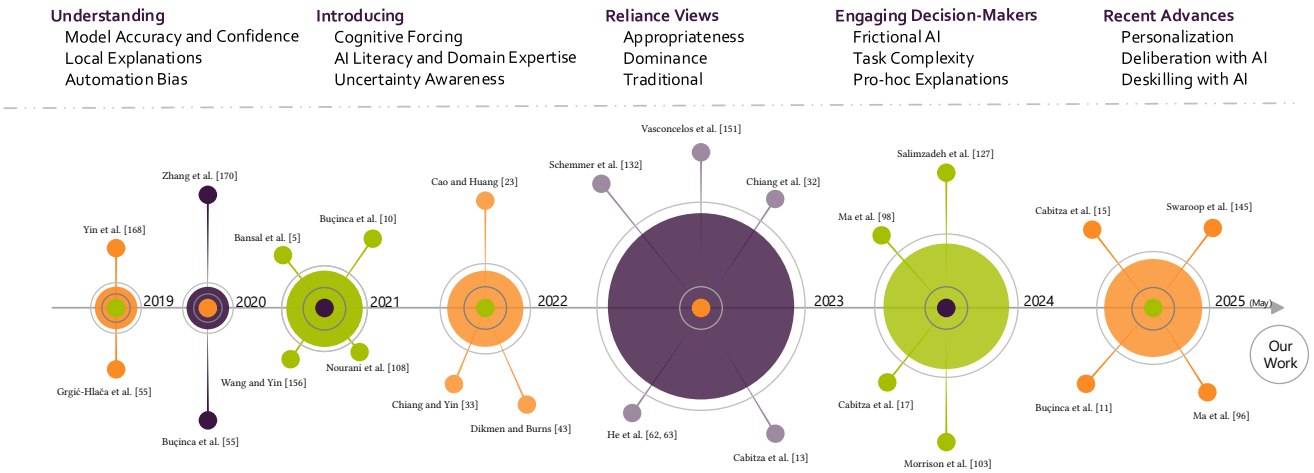

Figure 1: Timeline of key studies and concepts in human–AI reliance research. Pins indicate pivotal literature studies, and circle sizes represent publication volume per year. Over the years, research has shifted from understanding the effects of model-level features (local explanations, accuracy, etc.) to enhancing the capabilities of decision makers (end users), while defining methods and metrics to capture reliance. A critical debate exists in the literature between enhancing user engagement with AI systems (cognitive forcing, frictions, adaptations) and improving user expertise (domain knowledge, understanding).

Abstract

AI systems are increasingly being positioned to assist people in decision-making. However, recent empirical studies show critical concerns that people over-rely on AI advice without analytically engaging with it. While HCI research explores how people rely on AI advice, we argue that it largely overlooks an important aspect: replicating realistic decision-making scenarios. Human-AI interaction factors influence people's reliance on AI advice. To understand human-AI interaction factors and their interplay, we conducted an analytical review of recent studies in human-AI reliance literature. We analyzed the decision-making tasks in research and their validity in application-grounded contexts. Our findings show that user engagement is a precious commodity for relying on AI advice; however, it comes at a cost. We also discuss factors contributing to *"appropriate reliance"*, existing research gaps, and recommendations for intervention design for human-AI reliance. Our work contributes to the critical body of research on building appropriate reliance on AI advice.

CCS Concepts

• **Human-centered computing** → **HCI design and evaluation methods**; *Empirical studies in HCI*; User studies.

Keywords

Human-AI Reliance, Appropriate Reliance, AI Reliance, Human-AI Decision Making

ACM Reference Format:
Muhammad Raees, Vassilis-Javed Khan, Ioanna Lykourentzou, and Konstantinos Papangelis. 2026. Do People Appropriately Rely on AI-Advice? An Analytical Review of HCI Research on Human-AI Decision-Making.

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791467

In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3772318.3791467

1 Introduction

Artificial Intelligence (AI) systems continue to be *fallible* [77]. Although in many cases, AI performance is superior to human performance [5, 99], it still suffers from systematic or random errors. In such situations, tasks cannot be delegated *entirely* to AI, for reasons that include the need to ensure accountability or ownership over the decisions being made [13, 81]. Hence, researchers and practitioners posit AI to *assist* humans in their decision-making [5, 54], to enhance what the literature refers to as human-AI *complementary team performance* (CTP). Human-AI CTP is defined as the situation where humans and AI systems collaboratively achieve a performance superior to their individual performances [41, 134, 169]. Typically, in human-AI decision-making contexts, the AI system provides advice, and users make the final decision by accepting or rejecting/correcting this advice [13, 132]. Research [5, 60, 151] has explored methods to understand and enhance human-AI decision-making in diverse domains, including medical [17, 20, 71], finance [21, 28, 43], and justice [32, 52, 86]. The main objective is to achieve human-AI CTP, such that the users should only accept advice from AI when it is correct and reject advice when it is incorrect, coined as *"appropriate reliance"* [132]. A snapshot of research on human-AI reliance is depicted in Figure 1.

Building *appropriate reliance* is a complex problem, and no *panacea* can be applied to all situations [49, 81, 132]. There is mounting evidence that human-AI CTP often performs worse than AI systems alone [54, 83]. However, legal and ethical concerns with high-stakes tasks necessitate AI to support, instead of *automating* human decision-making [12, 147]. Research increasingly highlights the need to gain a foundational understanding of appropriate reliance on AI advice [13, 41, 81, 132] to improve human-AI decision-making scenarios. This understanding informs developers and stakeholders with a groundwork to implement AI support to assist humans in decision-making. Hence, it is important to understand the process through which decision-makers determine whether to rely on AI or not. This process should ensure that AI assistance is indeed complementary and enhances the decision accuracy, along with empowering users to rely appropriately, instead of *averting* [42, 143] or *complying* [91, 111] with AI blindly.

Research has explored several approaches for studying human-AI reliance, mainly categorized under enhancing the **AI systems** (explanations, performance, uncertainty) [5, 54, 118, 127], **understanding users** (biases, expertise, self-assessment) [48, 108, 126, 144, 151], and **designing interaction** methods [9, 10, 15, 60]. Each category in itself contains numerous fragmented theories, methods, and application areas that study *appropriate reliance* from different angles [10, 13, 103]. Recognizing these dimensions, we find that there is limited research that analytically evaluates recent HCI research in human-AI appropriate reliance. In this work, we analyze and consolidate complementary research approaches in human-AI *appropriate reliance*. Previously, Lai et al. [81] and Eckhardt et al. [49] explored empirical human-AI decision-making research, without an explicit focus on appropriate reliance. Contrary to existing works [49, 81], we address nascent challenges and the recent body of research, which has increasingly started to explore human-AI reliance through objective metrics. We analyze research studies around users, experimental designs, evaluation metrics, explanation methods, and implementation details. By doing so, we aim to provide recommendations towards building appropriate reliance and calling research focus on open challenges. Our research is driven by the following questions:

(1) How does current research *define and measure* appropriate reliance on AI advice?
(2) How can the state-of-the-art help researchers and practitioners understand and evaluate appropriate reliance on AI advice, focusing on the users, AI systems, and interaction methods it employs?

Our findings show that HCI research aims to improve human-AI reliance from different aspects to understand the users' *mental models* (including how they build understanding of the AI models, and how and whether they become aware their own confidence or uncertainty in interacting with AI systems), as well as *interaction methods* (including exploratory interfaces, frictional design, and engagement strategies). In terms of applications, studies employ both *real and hand-crafted* (simulated) AI models, potentially *introducing bias* and not capturing realistic AI behavior, mostly with *novice users* (as opposed to users that are a representative sample of the study's target population). *Realistic tasks* such as medical diagnostics and financial investments are explored with *non-expert crowd workers*, the involvement of whom does not necessarily replicate real-world experiences. Studies employ *objective* and *subjective* measures to capture user behavior and perceptions, based on research questions and tasks. We find a variational use of explanation methods and user engagement to reduce direct/default AI acceptance. At the same time, limited focus has been placed on studying users' interaction and *agency* over AI mechanics to contest and adapt decisions for improving reliance on AI. In a nutshell, our work makes the following contributions towards the goal of understanding research on achieving appropriate reliance on AI advice:

(1) We analyze recent empirical human-AI reliance research from January 2018 to May 2025.
(2) We discuss and argue on the *definition* of appropriate reliance, and advocate for research consensus.
(3) We derive and discuss the dimensions of users, AI systems, and interaction methods in studying human-AI appropriate reliance. We extensively analyze current research along those directions and elicit recommendations to enhance future studies.

2 Background and Research Position

Traditionally, AI systems have been *black-boxes* and *accuracy-driven*, being complex to be understood by their end users [2, 56]. Human-Centered AI (HCAI) research [137, 138] focuses on making AI systems more accessible, transparent, trustworthy, and accountable to end users [165], with the aspiration to eventually achieve complementarity [20, 82, 83, 117, 169]. However, for various reasons [20, 81–83], achieving human-AI complementarity is a hard problem [54]. Hence, HCAI research aims to enhance the efficacy of decision-making and to build reliance on AI systems through various intermediary steps, such as explanations, interactions, or trust

formation [56, 121, 137, 138, 166]. In the following sections, we provide a background on human-AI reliance and its progression, along with automation [143] and socio-technical systems' maturity [128]. We evaluate the literature focused on enhancing human-AI reliance using objective measures, and review the closely related works to further the research within the broader HCI landscape.

2.1 Trust and Reliance

Building reliance on AI advice is rooted in building trust [85, 130, 139]. The concept of trust has evolved over the years, from trust in automation (TiA) [65, 87] to trust in modern-day AI assistance [34, 74, 89, 154]. To understand trust, theoretical frameworks from psychology are used in various studies [4, 59, 87, 149, 150]. Using these frameworks, AI-assisted decision-making operationalizes *trust* and *reliance* measures, where trust reflects the users' subjective perception (i.e., affective or emotional attitude) towards others or systems [87, 100], while reliance pertains to their (observable) objective behavior [47, 130]. Initial works [65] mostly studied users' trust in automation; however, trust measures do not necessarily enhance reliance [131], which is gaining attention in recent work [81, 130].

To study reliance, research uses an approach called *Judge-Advisor-System (JAS)* [140], where a human acts as judge and gets advice from AI. To form reliance, this approach has to solve two main challenges: 1) the advice/advisor should be credible, and 2) humans should trust and then rely on the advice/advisor *appropriately* [132]. Recently, *"appropriateness"* of reliance has been defined to differentiate it as a *construct* in human-AI decision-making [131, 132]. For instance, Schemmer et al. [131] define the appropriate reliance as *"humans' ability to discriminate correct and incorrect AI advice and to act upon that discrimination."* Hence, discriminating between correct and incorrect advice is not sufficient, but humans also need to adapt their decisions accordingly to enhance human-AI CTP [67]. Mere discrimination can help detect errors in AI advice, but it does not enable contesting/correcting them. In addition, it is argued that users should neither *over-rely* nor *under-rely* on the AI advice [132]. *Over-reliance* is when users accept incorrect advice, and *under-reliance* is when users reject correct advice. Hence, we investigate how research explores the users' appropriate reliance, over-reliance, and under-reliance on AI advice.

Users can over-rely or under-rely on AI advice due to various reasons. For instance, users may over-rely on the system's advice, demonstrating *complacency* or *automation bias*, due to their belief in the system's superiority [73, 91, 111, 161]. Conversely, users may refrain from using the AI system due to *performance bias*, perceptions of *self-efficacy* (the ability to make decisions without advice), or purposefully considering the system's advice as inferior, demonstrating *algorithmic aversion* [42, 143]. The users' inability to understand the system operations and uncertainty, either due to the system's complexity or their lack of expertise, also affects reliance [2, 143]. To investigate such issues, research has explored understanding system operations [56, 111], and providing users with various forms of assistance, including explanations [50], uncertainty estimation [118], and interactions [45, 121].

Enhancing the explainability of systems supports users in understanding those systems in the first place [155, 165]. The literature on human-centered explainable AI (XAI) either advocates for transparent models or enhancing the interpretations of opaque models [50, 56]. XAI methods help users understand and augment their decision-making with AI, which has a direct impact on the formation of their trust. Similar principles are also applied to building reliance on AI advice, where users form a perception of AI performance based on its behavior [81, 100]. However, explanations of AI systems are often not enough to make users rely effectively on AI assistance [5, 129]. Hence, the combined study of *human-related* aspects (e.g., biases) [42, 148], *attributes* of the AI systems (e.g., XAI) [90, 94, 124], and *characteristics* of the decision-making tasks [54, 126] is essential to understand appropriate reliance.

2.2 Related Research Analysis

Existing reviews in HCAI can be broadly categorized into XAI [2, 6, 92], trust [100, 162], interactions (i.e., interactive Machine Learning (ML)) [45, 121], or related domains [1, 7, 49, 81]. For example, Bertrand et al. [6] reviewed the literature on how users' behavior with explanations affects AI-assisted decision-making. They examined the role of *XAI* systems on the users' cognitive biases and trust. Their research focuses on enhancing user acceptance of the system by improving explanations. Studying *trust calibration*, Wischnewski et al. [162] evaluated studies on the calibration of trust in automated systems up to 2022. This work [162] analyzes trust calibration and trust measurement in decision-making scenarios that include delegating tasks entirely to the system, and mainly focuses on autonomous and robotic systems. In a similar vein, Mehrotra et al. [100] surveyed research from 2012-2022, synthesizing several perspectives on forming trust with automation and AI systems. Their work looks into historical perspectives of trust and investigates disagreements on definitions and measures.

Understanding can also be improved through *interactions* [45, 121]. Raees et al. [121] review user interactions with AI assistance to enhance system acceptance. Their work mainly looks at XAI and proposes interactions to enhance user trust and agency with the system. In related domains, cognitive biases are also studied, which affect the users' decision-making with AI assistance. Boonprakong et al. [7] conducted a review that focuses on how researchers study various types of cognitive biases in automated/computing systems. Their work highlights the significance of designing AI systems that support human understanding and discourage biases. In the human-AI decision-making context, Lai et al. [81] analyzed empirical research outlining the progress and gaps for human-AI complementarity. Their work captures empirical studies in human-AI decision-making up to 2021. Eckhardt et al. [49] surveyed human-AI reliance research by employing a socio-technical lens. While these reviews highlight the needs and contexts for understanding trust and reliance, there is limited analysis of HCI research on **recent studies** that examine human-AI appropriate reliance. In the nascent field of human-AI reliance, an in-depth synthesis of the state-of-the-art is essential. The recent growth in human-AI reliance studies (Figure 1) also shows the significance of AI assistance for a wide range of stakeholders. Hence, in this work, we build upon the existing knowledge to augment the discussion on **human-AI reliance in HCI research.**

Our Contribution. Human-AI complementarity is highly important for AI adoption in real-world contexts. Despite recent

growth, limited work has been dedicated to reviewing studies on human-AI appropriate reliance in HCI research, which is scattered around trust and reliance [49, 81, 132]. As research increasingly focuses on *objective measures* of reliance [13, 132], it is important to analytically evaluate recent studies exploring *appropriate reliance* on AI advice. In this work, we use a standardized and systematic method to analyze studies by specifically focusing on research in human-AI **appropriate reliance** and covering **recent work**. Through our analysis, we summarize the current landscape (concepts and methods) in human-AI reliance research, identify nascent challenges, and present recommendations that support enhancing appropriate reliance on AI. We provide an analytical evaluation of current work from the lens of *users*, *AI systems*, and *interaction methods*. We believe this work will provide an in-depth synthesis for HCI researchers investigating human-AI decision-making contexts.

3 Methodology

In what follows, we explain our review protocol and criteria. To ensure high-quality reporting, we used a standardized protocol for the search and selection of relevant studies, as well as for the presentation of the results. We developed this protocol following recent guidelines on literature review research [141], and according to the PRISMA [110] framework, which is outlined in Figure 2.

3.1 Search and Identification

We identified queries with an initial search on the ACM Digital Library (DL) with keywords: *"human-AI reliance"*, *"appropriate reliance"*, *"over-reliance on AI"*. Analyzing article abstracts and our understanding of literature in *"human-AI reliance"*, we formulated our full search queries, as shown in Table 1. Based on our study's focus and the existing works [81, 132] highlighting nascent explorations in human-AI reliance, a timeline from Jan 2018 to May 2025 is used. We conducted an initial search using the SCOPUS database. To enhance the results, we conducted a similar search on the ACM Digital Library (DL). We collected 534 search results across the queried databases.

3.2 Selection Criteria

The initial screening consisted of a title and abstract assessment with soft inclusion criteria as follows:

- **Peer-Reviewed:** Peer-reviewed full articles (conference or journal). Short papers, workshop papers, books, chapters, editorials, and non-reviewed work (e.g., from arXiv) were excluded.
- **Primary Studies:** Only primary studies, excluding secondary studies such as reviews, surveys, and similar.
- **Human-AI CTP:** Articles that study decision-making aspects, such as through assistive AI. Articles that lacked focus on human-AI CTP or only compared user behavior on experts' (human) advice were excluded.

This screening reduced the total number of articles to 73. We subsequently applied a more comprehensive set of inclusion and exclusion criteria to these articles by reviewing their full texts, as follows:

(1) **Conceptual or Unusable:** Articles that only studied conceptual or algorithmic models for human-AI complementarity (i.e., reliance on technology usage) were excluded.
(2) **Human-AI Decision-Making:** Articles that studied human-AI assistance, but did not focus on or measure objective user behavior for reliance measures [69] (i.e., only capturing users' trust perception/feedback [124, 158]) were excluded.
(3) **Building AI Reliance:** Only the articles that study building and measuring reliance on AI assistance using objective measures were included.

Following the aforementioned criteria, two researchers independently reviewed each article and coded them in categories from 1 (*conceptual or unusable*) to 3 (*Building AI Reliance*). Inter-rater reliability using Cohen's Kappa [84] indicated strong agreement ($k = 0.80$). Conflicts in the codes were discussed to form a consensus. This process resulted in 41 primary studies, with category 3 being included for analysis.

To further expand our selection, we performed iterative backward and forward (using Google Scholar) **snowballing** [163] by analyzing references and citations of each paper. We used the same inclusion and exclusion criteria on the snowballing results and selected an additional 15 of 20 identified studies (Backward: 10; Forward: 5). This provided us with a total of 56 studies for final analysis. The review protocol and process are provided as an Open Science Framework (OSF) repository [3] for the reader's evaluation.

3.3 Bibliometrics

We first performed a meta-analysis of the selected studies, coding each paper for domains, tasks, evaluation methods, and AI methods used for decision-making. Then, we analyzed the studies focusing on similar patterns and introducing new concepts. The categorization of the selected studies along these dimensions is expanded in Section 4 and summarized in the appendices. Figure 3 shows a meta-analysis of the dataset based on the year and venue[1] of publications. A recent surge in publication volume reflects the focus of the HCI communities on studying appropriate reliance, transitioning from existing methods that only focused on studying subjective trust and user confidence. The venues of the selected studies are quite diverse; however, the majority of the studies were published in reputed HCI venues such as PACMHCI (30%), IUI (25%), CHI (23%), FACCT (5%), JAI (4%), and others (combined 13%).

3.4 Reliance Views and Concepts

"Objective reliance" is defined as a behavioral aspect with three patterns/views [26, 131], which are expressed in Table 2. The *traditional view* [151] assesses whether the user over-relies or under-relies on AI by studying their behavioral patterns. This view, with some additional measures, is studied more commonly than the others in the identified corpus, without explicit definitions. Schemmer et al. [132] define the *appropriateness view* as attempting to understand the users' Relative Self-Reliance and Relative AI Reliance.

[1] PACMHCI: Proceedings of the ACM on Human-Computer Interaction. CHI: Conference on Human Factors in Computing Systems. IUI: Conference on Intelligent User Interfaces. FACCT: ACM Conference on Fairness, Accountability, and Transparency, JAI: Journal of Artificial Intelligence.

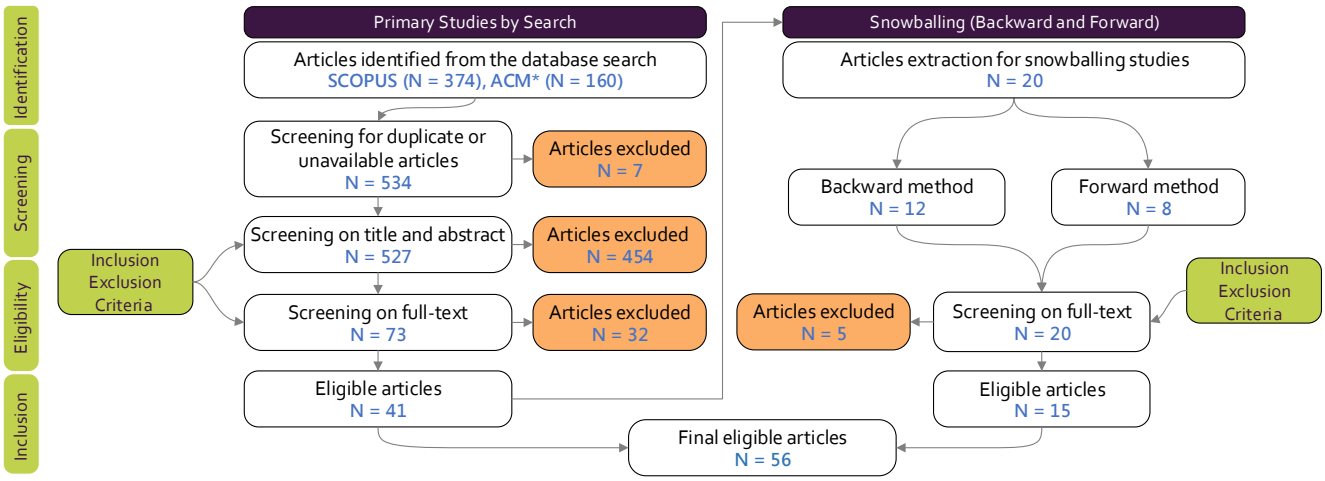

Figure 2: Overview of the analysis framework. The primary search was conducted on SCOPUS, with a secondary search on the ACM Digital Library (DL). The SCOPUS search comprises articles from diverse sources, while the ACM DL results focused primarily on HCI venues. *The ACM results were restricted due to overlaps with the retrieved corpus.

Table 1: Queries selected to conduct database search (SCOPUS and ACM DL). The search queries were adapted to the interface of each database and complemented with known abbreviations (AI/ML) where needed. The search queries ensured that the articles were selected from broadly representative human-AI domains.

Search Queries
("over-reliance" OR "under-reliance" OR "reliance" OR "appropriate reliance" OR "human-AI reliance" OR "trust") AND ("artificial intelligence" OR "machine learning") AND ("decision making" OR "decision support" OR "human-AI complementary performance", OR "human-AI interaction" OR "AI assistance" OR "user behavior")

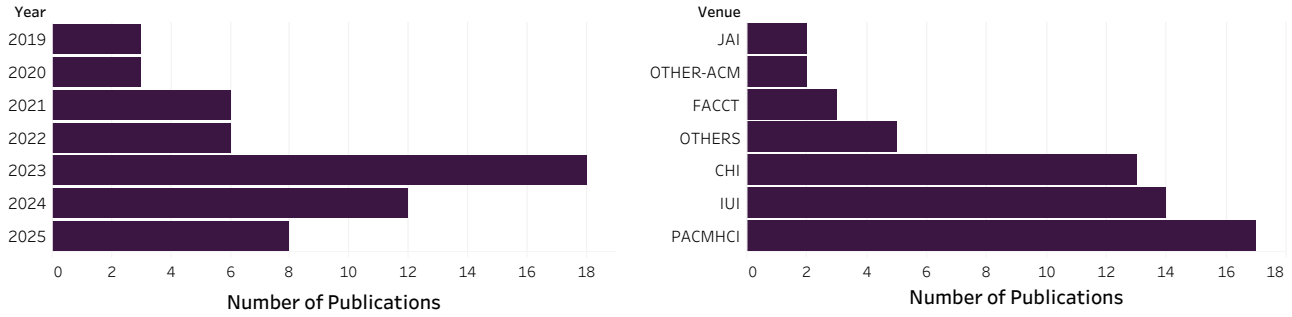

Figure 3: Year-wise and venue-wise distribution of the selected articles, showing a growing interest in human-AI reliance within HCI research over time. The year 2023 represents the highest number of selected articles. Venue-wise, the majority of the articles come from PACMHCI. Venues with a single article are grouped as "OTHER-ACM" (within ACM) and "OTHERS" (outside ACM). Approximately two-thirds of the selected articles (38 articles) were published between 2023 and 2025.

Relative Self-Reliance (RSR) is a measure that describes the situation in which humans correctly reject wrong AI advice and make the correct decision independently (i.e., without being influenced by the AI's wrong recommendation). Relative AI-Reliance (RAIR) captures the user's shift to a correct AI advice (i.e., the situation in which the user initially makes an incorrect decision but subsequently changes it to align with the correct AI recommendation). Cabitza et al. [13] define the *dominance view*, which captures the dominance that technology exerts on users, which can be detrimental or beneficial in overall decision quality. This view is often

Table 2: Patterns established in research exploring reliance in human-AI collaboration. The iterative decision-making pattern asks users to make decisions before and after AI assistance. 1: correct advice/decision or 0: incorrect advice/decision w.r.t. the ground truth. Ensuring correct/appropriate/beneficial reliance is the main objective of AI-assisted decision-making. The *traditional view* uses over-reliance and under-reliance measures based on behavioral patterns. The *appropriateness view* uses relative self-reliance (RSR) and relative AI reliance (RAIR) to measure users' independent and AI-dependent decisions. The *dominance view* measures the benevolent or detrimental exertion of technology for final decisions.

Human Decision	AI Advice	Final Decision	Traditional View [151]	Appropriateness View [132]	Dominance View [13]
0	0	0	Over-Reliance	N/A	Detrimental Reliance
0	0	1	Appropriate Reliance	N/A	Beneficial Under-Reliance
0	1	0	Under-Reliance	Incorrect Self-Reliance	Detrimental Self-Reliance
0	1	1	Appropriate Reliance	Correct AI Reliance	Beneficial Over-reliance
1	0	0	Over-Reliance	Incorrect AI Reliance	Detrimental Over-Reliance
1	0	1	Appropriate Reliance	Correct Self-Reliance	Beneficial Self-Reliance
1	1	0	Under-Reliance	N/A	Detrimental Under-Reliance
1	1	1	Appropriate Reliance	N/A	Beneficial Reliance

studied in high-stakes domains (e.g., medicine) to understand the beneficial and detrimental effects of using AI advice.

Morrison et al. [103] slightly expand the appropriateness view from Schemmer et al. [132], including the XAI dimension, i.e., where AI predictions can be correct, but their explanations can be incorrect or vice versa. Various methods use other measures, such as user agreeableness, or switching fractions with AI advice [13, 132, 151], as elaborated in Section 4. However, in all views, the burden of making a decision lies on the human user in the end. Several studies, however, do not explicitly conform to capturing these views but use similar measures for studying reliance. Regardless of views and slight variations in the definitions, research highly focuses on understanding human-AI reliance [124, 132]. Hence, in a nutshell, to form appropriate reliance, users must switch to the AI advice when it is correct and override it when it is incorrect.

4 Analysis

In what follows, we provide an analytical evaluation of the selected studies. Table 3 provides an overview of the main dimensions of our analysis. Section 4.1 examines the main research directions of the literature. We analyze the measures (objective and subjective) and measurement protocols used in the literature in Section 4.2. The analysis of the studies' experimental design is presented in Section 4.3. Finally, Section 4.4 explores the intervention used by the studies in our corpus and their impact on achieving appropriate reliance. The Appendices (Tables 5 – 8) provide additional information regarding our analysis, including a quantitative overview and detailed coding of the selected studies.

4.1 Studying Appropriate Reliance

Broadly, studies have focused on **three main directions** for enhancing appropriate reliance on AI, namely 1) **AI systems** (e.g., by improving performance, uncertainty, explanations) [5, 9, 118, 127, 156, 168, 170], 2) **user factors** (e.g., focusing on expertise, cognitive biases, traits) [33, 63, 98, 108, 123, 144, 145, 151, 153], and 3) **interaction factors** (e.g., focusing on enhancing engagement, or changing the behavior of the user towards the tasks) [10, 13, 15, 40]. Figure 4 illustrates an overall categorization of these three directions that the evaluated AI-reliance literature has focused on, and their sub-categories.

4.1.1 AI System Factors. Around 64% (36) of the studies specifically focus on **enhancing the AI system**, for example, by improving its performance or augmenting the explanations regarding its functionality (XAI), to support user reliance. Approximately 13 (23%) of the studies use AI performance metrics like model confidence and uncertainty [27, 28, 98, 118], as well as accuracy [168, 170] to understand user behavior. For example, Zhang et al. [170] study the impact of *model confidence*, which helps calibrate people's trust in AI models. Making users aware of system capability [62, 98, 168, 170] and uncertainty [24, 76, 118] is important; this is achieved by, for instance, conveying model accuracy on the dataset. Communicating about uncertainty in the AI system's capacities can help users gauge the pitfalls of model performance and associated risks. However, it can also give them a *false impression* of high confidence. Specifically, accuracy metrics can impact people's trust and reliance, with users reporting higher *trust* in the model when they can compare *their own* accuracy to that of the *system* [62].

With the majority of the studies using XAI, we find that around 57% (32) of them focus on **adapting the XAI** method used. These studies use (experiment) conditions to test the effect of feature-based [60, 153, 156, 164], example-based [17, 23, 108, 146], model-based [8, 33, 55, 114], or general explanations methods [102, 117, 127]. In addition, studies attempt to adapt explanation methods by changing their medium (text, visual) [103, 167], complexity (easy to complex) [151], context (analogies, reasoning styles) [27, 62], or model confidence [145]. For instance, model-based explanations may be used if the model is sensitive to the dataset [170], explanations adapted to AI's confidence may be used if the model is focused on prediction [5], or, if the model is focused on user performance, the explanations may be adapted in terms of *when* they are shown to the user [144].

Approximately two-thirds of the studies (38) use *local model explanations* [133, 153, 170] as compared to (13 studies) *global explanations* [33, 55, 62, 168]. Meanwhile, 22 (non-exclusive) studies use *example-based* [23, 23, 131, 167] explanations, while a few use

Table 3: Categorization and summary of main dimensions for the analysis. Dimensions include three main research directions explored in the literature to study appropriate reliance. The analysis also includes other dimensions, such as metics, measurement protocols, applications, participant expertise, recruitment, workload, remuneration, and AI system fidelity. We discuss common interventions and their impact on appropriate reliance. Sub-categories also include the number of studies in which they are explored to provide a quantitative summary of the analysis.

Dimensions	Categories	Sub-Categories (# Studies)	Explanation
Main Research Directions	AI Systems	XAI Adaptations (32) AI Performance (13)	§4.1 explores human reliance on AI by analyzing 1) adaptations of XAI and AI performance metrics, 2) impact of user factors (e.g., cognitive biases) and self-assessment, and 3) interaction with the AI system or motivating users through incentives, or providing performance feedback.
	User Factors	Biases (13) Self-Assessment (10)	
	Interaction Factors	Adapting Interaction (10) Motivating Users (29) Performance Feedback (9)	
Measuring Human Reliance on AI	Metrics	Objective Metrics (56) Subjective Metrics (51)	§4.2 examines objective and subjective metrics reported in the literature, along with their usage, in concurrent or multi-step decision-making protocols.
	Measurement	Concurrent (Single Step) (20) Sequential (Multi-Step) (40)	
Experimental Design	Domains	Healthcare/Medicine (10) Education/Learning (17) Business/Work-Context (20) Leisure/Sports/Arts (11)	§4.3 explores the contexts of applications (e.g., domain areas, implementation details), and participant expertise (e.g., background, domain knowledge), recruitment method/category (e.g., crowd workers), participant workload (tasks) and payment, and the level of AI implementation fidelity (realism, complexity, models, datasets, etc.).
	Participants	AI Experts (2) Domain End Users (6) Novices (48)	
	Recruitment	Crowd-workers (40) Non-Crowd-workers (16)	
	Workload	Crowd-workers (40) Non-Crowd-workers (16)	
	AI Fidelity	Actual (37) Simulated (19)	
Reliance Intervention	Impact	Positive (13) Negative (12) No (Explicit) Effect (31)	§4.4 explores the prospective themes and the impact of common interventions on achieving appropriate reliance.

* Sub-categories (studies) are often non-exclusive.

counterfactual or contrastive explanations [11, 17, 88]. Counterfactuals focus on the distinctions between the AI's suggestion and a likely human response while highlighting only the dimensions in which the two choices differ [11]. Buçinca et al. [11] study the benefits of contrastive explanations to enhance human skill on the task (i.e., human learning), comparing with unilateral explanations. To reduce the bias towards one advice or explanation, Cabitza et al. [17] use *"pro-hoc"* explanations (examples and counterfactuals) as a way to present multiple answers instead of AI advice to reduce inappropriate reliance.

Inappropriate reliance can also occur in the presence of **imperfect XAI** [103], a phenomenon where an explanation reveals evidence that does not necessarily comply with the prediction. Some studies [27, 62] adapt reasoning (analogies, inductive, deductive) of explanations, which show mixed effects on the reliance. Overall, these studies show how AI systems are adapted to users; however, there are various trade-offs with each method, and when complemented with user factors, the AI systems should be carefully evaluated to formulate what information should be explained.

The debate whether explanations increase or decrease reliance is ongoing [5, 10, 40, 151]. Vasconcelos et al. [151] discussed that the cost of engaging with explanations (i.e., user difficulty with XAI) can determine the user's reliance. For example, if understanding the explanation is equally or more complex than understanding the AI decision, users might refrain from it. Schoefer et al. [133] found that if explanations highlight the task-relevant features, it impacts users to engage with AI recommendations more closely. Over-reliance due to explanations can be calibrated with system confidence and uncertainty communication [151].

4.1.2 User Factors. Users have different biases towards automation and AI systems, which come into play when relying on AI advice. We find that around 23% (13) of studies explore various biases, including *automation bias* [108, 153] and *algorithmic aversion* [13, 112]. For instance, users may form a misbelief that they understand how the model works when they do not [5, 108]. Users may also develop an aversion to AI, being skeptical of the system due to their negative first impressions [108], thus under-relying on its advice. In addition, users might not be motivated enough to engage with AI advice, and thereby accept any system advice that comes their way [151].

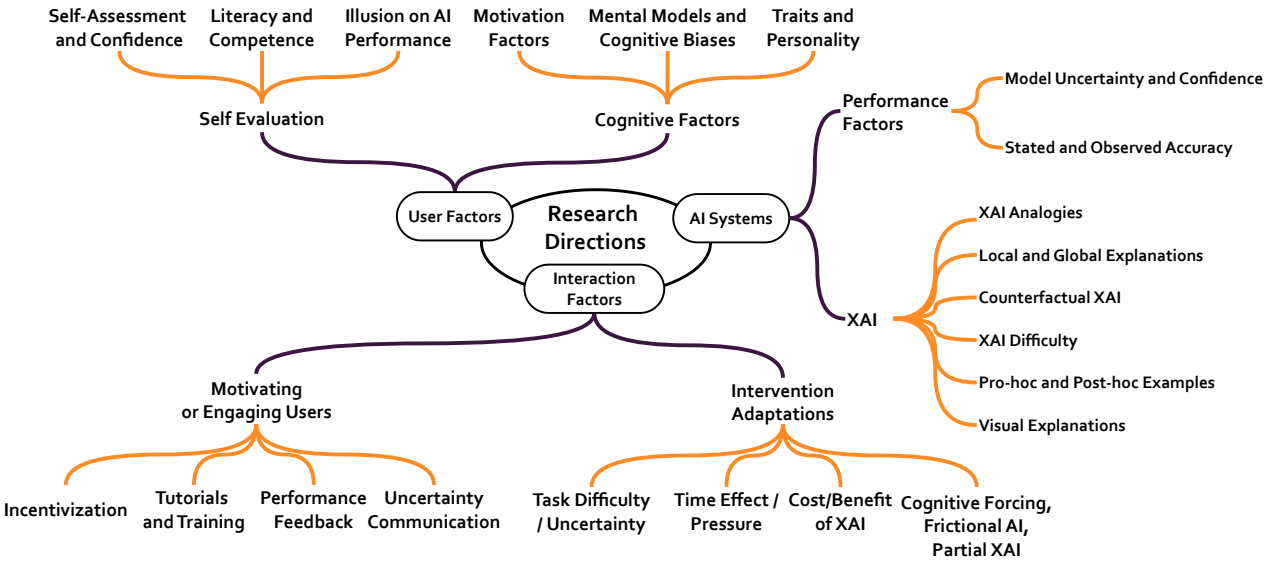

Figure 4: Overview of the methods and concepts used for appropriate reliance adopted in the final corpus. Studies adapt AI systems, such as their performance and explanations, to influence users to engage with AI advice. This includes understanding the effects of model performance and XAI methods on users' reliance behavior. Research also explores understanding user cognitive factors and self-assessments measuring users' mental models, biases, personality, expertise, and confidence. In addition, research investigates how adapting AI interaction through various engagement-enhancing methods, such as cognitive forcing or motivating/enabling users to perform better and build appropriate reliance.

Prior studies [54] in human-AI decision-making have shown that **cognitive biases** affect people's interaction with AI advice. For example, the order of observing accurate and inaccurate predictions affects the user's mental model [108]. Research aims to understand the users' mental models [18, 98, 118] to better predict their behavior, and subsequently design interventions. An important trait for improving the users' mental models concerning AI is capturing their **"Need for Cognition"** [19], a stable personality trait that captures one's motivation to engage in effortful mental activities [10]. Exploring mental models, AI can be *delegated* to solve easier problems, and users can be engaged for difficult ones, for instance, to recognize AI errors [5, 8].

Aside from mental models, other mediators affecting reliance on AI include **personality** traits [144, 145], **intuition** [30], and self-perceived **expertise**, which can all influence decision-making [62, 107]. We found that around 18% (10) of studies model and try to calibrate users' self-evaluation, such as their perceived degree of expertise [43], awareness [63, 97, 98], confidence [71], literacy [33], and their uncertainty [118]. While self-assessing, users can also create an *illusion of AI performance* [63] that can downgrade their performance as people often consider their judgment as a reflection of reality (**"naive realism"** [125]) without knowing their limitations, leading to over-reliance on models [94]. As expertise is an essential criterion to influence biases and reliance on the AI advice [24], Zhang et al. [170] try to enhance users' expertise by providing additional information. Dikmen and Burns [43] provide **domain knowledge**, such as contextual data and statistics, to improve expertise, but having no real incentives or the risk of making wrong decisions affects users' behavior.

A few studies [33, 61, 63] looked at providing XAI **tutorials** to convey AI performance and enhance users' abilities. Chiang and Yin [33] show that users need to understand and learn ML systems before they can effectively engage with them, allowing users to construct their datasets and tests to improve their mental models. Tests can identify possible disparities in data, and users can learn about these disparities before experimenting with tasks [29]. Users also face inherent issues with the complexity of explanations. For example, feature-based explanations (summing up feature effects) to show model confidence are not obvious for people without sufficient AI expertise to understand the underlying models [170]. Users may find those explanations to be rather foreign and mentally taxing to consume [156].

Users could trust an AI model *inappropriately*, and still achieve a higher accuracy (e.g., blindly trust a model that has a higher accuracy than oneself) [156]. Conversely, users can overestimate their confidence in decisions due to inaccurate interpretations [94], or cognitive biases (e.g., Dunning–Kruger effect [79]: people overestimate their abilities), leading to under-reliance on the ML models. For non-expert users, gaps in domain knowledge can lead to misunderstandings [43], which can subsequently affect the collaboration strategy [5]. Hence, understanding user factors is critical to conducting human-AI experiments and designing interventions to better adapt interactions with AI systems.

4.1.3 Interaction Factors. Most of the research on interaction factors focuses on adapting user interaction with the AI system, drawing from human psychology theories [36, 72, 159] and implementation design. For example, interaction can be designed to adapt (to) user behavior [5, 40, 145] or system functionality [9, 10]. We found that 70% (39) of the examined studies adapt the interaction through intervention methods (10 studies) or motivating factors (29 studies). A common method is using cognitive forcing functions (CFFs) [10, 40, 71] or frictional AI [13, 15, 17] to **add delays** to user engagement with AI. People employ dual process theory [72, 159], quick (system 1) or analytical (system 2) thinking when interacting with tasks. Buçinca et al. [10] use CFFs to increase people's **cognitive motivation** for engaging analytically with AI explanations. CFFs are used in frictional design [36, 105], which add delays or slow downs to force users to use (system 2) analytical thinking [13, 15, 71]. In similar lines, de Jong et al. [40] employ **partial explanations** as a way to increase user engagement with XAI. Ma et al. [96] introduce **human-AI deliberation** as a way to discuss conflicting opinions between AI and humans to create evidence-based decision updates for AI or users. Combining with **slowdown interventions**, studies explore adapting the decision-making approaches, such as *multi-step decisions* [131, 146], evoking or overcoming users' biases towards or against the AI advice.

Studies also engage users to enhance their competence [8, 23, 33], such as through **training**, or by providing performance feedback [55, 168]. However, due to factors such as task expertise and users' relevance with the tasks, *training methods* have not contributed significantly towards reliance [127]. Some (9) studies provide performance feedback to users to reassure their decisions; however, others reason that this is not a viable representation of the real world, as decision feedback is not instant but comes in the future. In some cases, the feedback is provided during the training tasks, and not during the actual experiment. Another commonly explored method (29 studies) is to improve engagement by enhancing the users' incentives, although with a meager value. For instance, many studies [28, 60, 127, 144, 168] explore **extrinsic incentives**, such as monetary rewards, to improve performance; however the use of such incentives rarely achieves the intended outcomes [55, 164]. Hence, incentives, at the current rate, have not been very effective in enhancing users' reliance on AI advice.

Sandbox environments [33, 61] are used to evaluate users' engagement with model predictions, where *interactivity* or *agency* can enhance people's understanding and perceived usefulness of the system. Studies also add time pressure [22, 123, 144, 145] to understand engagement for reliance under constraints. Interaction methods are adapted to add task constraints such as complexity or uncertainty of tasks [127]. Engagement can also make people evaluate a model's predictions [117]; however, without the model's information, people's reliance depends on alignment with the model's prediction [94]. Overall, adapting interaction to users' performance and cognition, and AI's performance and confidence has been found useful [10, 13, 15, 40], but engaging users to elicit realistic behavior is still a challenging task.

Key insights: The literature has focused on three main directions to improve user reliance on AI. The first one concerns **AI Systems**, where studies have explored various XAI methods, XAI delivery mediums, and XAI complexity levels, without a conclusive consensus emerging on whether explanations increase or decrease reliance. Secondly, the literature has explored **user factors**, examining how human biases and other traits (e.g., personality, intuition, and self-perceptions) affect user reliance on AI. Notably, users' "need for cognition" can be used to predict whether users will meaningfully engage with AI. Third, the literature has looked into **interaction Factors**, with a notable finding that cognitive forcing and frictional designs show a positive impact towards users' reliance on AI; however, extrinsic incentives, such as monetary rewards, do not appear to contribute towards appropriate reliance.

4.2 Measuring Human Reliance on AI: Objective and Subjective Metrics and Measurement Protocols

We now move to the **metrics and protocols** that studies have employed. Figure 5 shows the spread of both objective measures for reliance and subjective measures for capturing the users' perceptions of AI systems.

Starting with the analysis of *objective metrics* and approaches to human-AI reliance, we find that **decision accuracy** [108, 132] is the most common metric (35 studies) to evaluate whether the user made the correct final decision or not. Other important metrics, which are often used together, are **agreement fraction** (21 studies) [62, 63] and **switch fraction** (15 studies) [60, 102]. Agreement fraction measures how frequently (or what percentage of decisions) users agree with AI advice, and switch fraction measures how frequently (or what percentage of decisions) users change their initial decision after seeing the AI advice. Neither metric, however, captures whether the reliance was appropriate, i.e., the correctness of user decisions. A few (4) studies use the *weight of advice (WOA)* [13, 33, 117] to measure if the users change their decision due to AI [132].

Slightly different than agreement and switch fractions, recent studies also report **over-reliance** [10] (12 studies) and **under-reliance** [27] (10 studies) as separate measures, which capture the correctness of decisions, measuring when users blindly agree with incorrect advice and neglect correct advice, respectively. Some studies also use **relative self-reliance** (RSR) (9 studies) and **relative AI reliance** (RAIR) (10 studies). As also explained in section 3.4, RSR measures the human ability to reject wrong AI advice and make a correct decision, while RAIR measures users' ability to correct their own (pre-AI) decision, following correct AI advice. To evaluate user interaction on these measures, 79% (44) of the studies use *empirical analysis*, while a few (11 studies) use *mixed methods*, and one study [118] mainly uses *qualitative analysis* (e.g., interviews). Additionally, think-aloud [9, 30] and survey feedback [8] are also used to provide deeper insights about human behavior in reliance strategies.

The point at which AI advice is given, i.e., using a **concurrent** or **sequential** (multi-step) protocol, can affect how users perceive this advice [5]. Concurrent (single-step) and multi-step processes slightly differ in the way they use reliance metrics [131]. For instance, concurrent approaches do not permit the user to make their own independent decision before receiving the AI advice, resulting in fewer objective metrics reported. In these single-step decision

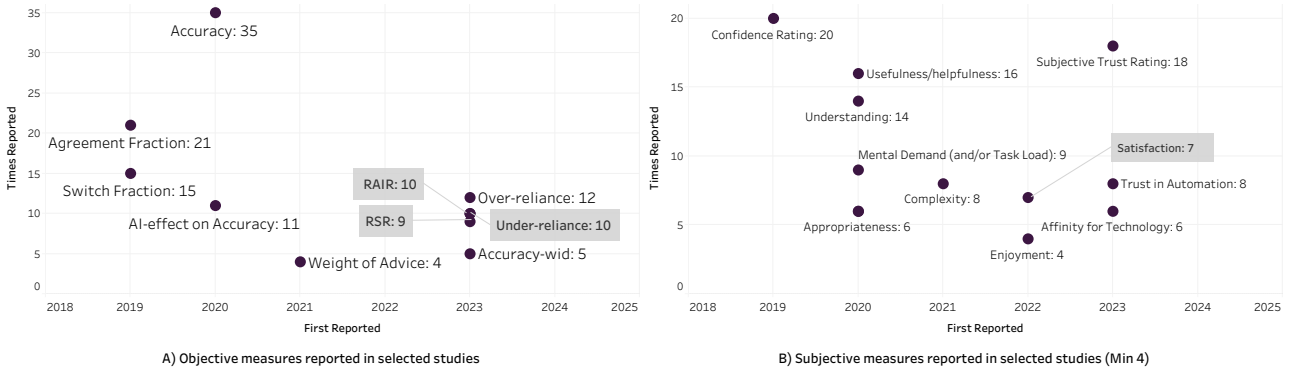

Figure 5: Left: The count of objective measures and the year first reported in selected studies. Accuracy, which can be used in both one-step and multi-step approaches, is the most widely used measure. Studies also measure users' reliance on AI through agreements or switches. AI's effect on accuracy and RAIR measure when users made a correct decision due to correct advice. Right: The count of subjective measures and the year first reported in selected studies. Studies commonly use confidence, trust, usefulness, and understanding ratings to measure subjective experiences. Some studies also capture users' mental demand, complexity, and enjoyment to understand experiences better.

studies, users are provided with AI advice and asked to decide without making initial decisions. Hence, observed agreements between the human and the AI, in the concurrent paradigm, could arise either because the human took the advice, or the human had already arrived at the same judgment independently of the AI [146]. Thus, the concurrent advice approach may induce biases towards AI assistance without analytically evaluating the pre-AI user decision first. Measuring agreement or switch fractions is impractical in a concurrent approach [146]. Therefore, 71% of the studies (40) mainly use a two-step (or multi-step) process where humans first make the decisions, investigating the task themselves, and then are provided with the AI advice, subsequently having the option to revise their decision. The two-step approach can reduce the direct over-reliance, but it can also induce an anchoring effect [132].

Most studies rely on fixed reliance metrics (i.e., for each interaction), and user **reliance development over time** is underexplored [22, 40, 145]. Other cases include reliance development over multi-stage decisions, where users explore additional information (in intermediary steps) before making decisions [40]. Different from others, Cao and Huang [23] explored an interesting direction to capture real-time behavior by analyzing the users' eye gaze to assess human-AI reliance. While this approach may indeed capture detailed insights across a timeline, using physiological measures is hard to replicate in most practical settings of human decision-making scenarios.

Studies often supplement the objective measures of human reliance on AI with various **subjective metrics**, aiming at capturing user perceptions towards AI. Commonly used subjective metrics include the users' perceived **trust** (Likert scales [102] (18 studies), or Trust in Automation [60] (8 studies)), **confidence** [40] (20 studies), **usefulness** [15] (16 studies), **understanding** of advice [112] (14 studies), the users' **mental demand** [10] (9 studies), **satisfaction** [97] (7 studies), **appropriateness** [11] (6 studies), and others [33, 127, 145]. Subjective measures focus on measuring the perceived usefulness and understanding of AI advice to capture whether users are comfortable with the system. These measures are mostly self-assessed attributes and are also used to understand the impact on user performance, but they do not necessarily help extract actual reliance behavior. A summary of the objective and subjective measures used by the examined literature is provided in (Appendix) Table 5.

Due to the variance in using objective criteria to measure users' performance, **there is still limited consensus on common measurements** of human reliance on AI across the studies. The combination of objective and subjective measures does not necessarily lead to better predictions of reliance; for instance, users can express high **trust** towards a system that aligns with them, without any objective **reliance**. While measuring reliance, user experience is often overlooked; however, it is important for evaluating user interaction with any new technology or automation – AI included. Finally, some studies [57, 134] measure reliance as merely the presence of an AI system (i.e., deciding with and without AI systems in separate conditions), which falls out of the scope of this work.

Key insights. The literature reports the use of a variety of objective and subjective measures to evaluate the users' reliance, as well as their perceptions towards AI. In terms of **objective measures**, we find decision accuracy, agreement, and switch fractions are commonly used. Recent work has defined tailored measures (e.g., RAIR, RSR) to capture users' over- and under-reliance. The protocol of providing AI advice is also crucial. Concurrent study designs, where AI advice is given before making independent decisions, may induce biases, as it may not be evident whether the user's decision is due to the AI advice or due to their own existing views. Multi-step study designs report more objective measures and allow for a more analytical evaluation of AI systems. In terms of **subjective measures**, we find several metrics like trust and confidence for gauging the users' perception of AI systems; however, subjective impressions (trust, confidence, etc.) do not necessarily reveal whether users can build appropriate reliance on AI.

4.3 Experimental Design

We now look deeper into the experimental design of the human-AI reliance studies, and more specifically, on application domains, participants (recruitment, the workload, and the remuneration), and AI implementation (fidelity).

4.3.1 Application Domains. Understanding user behavior towards AI also depends on the application and task domain. Studies have employed different applications in high and low-stakes domains, including medicine/healthcare, business, education, and leisure, among others. Table 6 (Appendix) elaborates on the wide range of **application domains** and decision tasks that have been used by the selected corpus.

4.3.2 Participant Expertise. Despite the real-world significance of the selected applications, most (48 out of 56) studies use **lay users**, most notably crowd workers (40 studies), as the study subjects performing the tasks. Only a handful of studies (6) employ **domain end users**, usually for medical diagnostics. While these human-grounded evaluations with crowd workers are appealing when assessing target users, they only capture the general notion of human-AI decision-making [44]. Involving crowd workers is an established practice in HCI research [44]; however, understanding domain users' decision-making perspectives is also important.

From all the studies in the evaluated corpus, **very few involved users with expertise in AI** (2 studies) [30, 167]. The inclusion of *actual/potential users* with relevant task expertise as the subjects of the study remains rare as well, outside of physicians for the medical domain [88, 164]. Studies try to enhance the stakes/relevance of the tasks by emphasizing their *realistic* nature and linking their motivation to real-world, such as medical diagnostics [13], recidivism [32], pricing houses [118], approving loans [60], predicting student performance [123], and other identification tasks [22, 94], highlighting the relevance of their implementations to actual scenarios (Table 6).

4.3.3 Recruitment. Study participants are recruited using different methods. Figure 6 shows the prevalence of different users and recruitment methods in the evaluated studies, where **novice crowd workers** (lay users) are recruited through platforms, such as Prolific and Amazon Mechanical Turk. While most studies use novice crowd workers, a few report that the hired crowd workers have some *domain expertise* (e.g., having served on a jury [55]). Apart from crowd workers, novice users (12 studies) are also recruited through universities (students) or via advertisements/emails, while domain experts are *mainly* recruited through (convenience) sampling (4 studies). A typical study recruits on average approximately 260 people (*mean = 260.85, std = 279.4*), while the recruitment is sometimes (23 studies, 41%) split across multiple experiments. **Studies conducted with crowd workers show a significant difference in participant count** (*mean = 343.35, std = 284.82*) compared to non-crowdsourcing channels (*mean=35.33, std=24.11*).

Around 54% (30) of the examined studies acknowledge the limitations of recruiting **laypeople** and using **tasks that are a proxy or not relevant** to the users' expertise or decision-making stakes. Specifically concerning task expertise, around 29 (52%) of the studies did not use tasks that were relevant to the users' expertise, e.g., recidivism [32, 55, 156], income/price prediction [30, 97, 98, 117, 170], loans/investment [43, 62], or medical decisions [24, 71, 76, 145] for laypeople/crowd workers. Chiang and Yin [33] note that real estate experts can have expertise that laypeople lack. Existing work [45, 122] also highlights that running complex tasks with crowd workers may not be entirely representative. Dikmen and Burns [43] note that non-expert users' expectations from AI are different from those of users with domain expertise.

Around 19 studies with laypeople use tasks that can be performed without domain expertise, such as reviews [5, 18, 93, 133], speed dating [94, 168], puzzles [23, 114, 151], or nutrition suggestions [9, 10]. Studies increase the stake by linking (bare) **incentives** to performance [55, 168] to capture the *"users' attention and engagement"* with tasks. Still, **the lack of contextual relevance reduces user engagement**, irrespective of monetary rewards. Buçinca et al. [9] argued that accessing domain expert users for experiments is notoriously challenging and costly [10]. However, **if the tasks lack real-world significance and involve no real-world stakes for participants, the study may fail to elicit realistic user behavior**, thereby limiting the validity of its findings concerning human reliance on AI [40].

4.3.4 Workload and Remuneration. Figure 7 maps the crowd workers' task volume and time spent on the task, and their corresponding payment levels (in USD, hourly rates). We observe a **negative correlation between task volume/time and user payment** (*mean = 10.99, std = 4.24, min = 1.5, max = 25*), raising some concerns on the rigor of intensive decision-making tasks. Workers perform considerable tasks (*mean = 27.56, std = 40.32*), while studies aim to enhance their engagement and critical thinking, proportionally in a short time (*mean = 20.68 mins, std = 8.61*). Studies with non-crowdsourced participants report relatively higher experiment time (*mean = 45.27 mins, std = 30.88*), while the payment, only where reported, is also relatively higher (*mean = 21.28, std = 28.0*).

He et al. [60] acknowledged that **crowd workers rush through the tasks** and provide low-effort results. Studies use attention checks; however, with decision-making tasks, that strategy alone does not mitigate the problem of low-effort results. Moreover, users who are unfamiliar with the study scenario doubt their ability to give the right advice or decisions [118]. Morrison et al. [102] also share the concern that complex tasks are difficult for crowd workers to engage properly. Hence, further work is required to ensure that different types of users (with varying AI and domain expertise and/or other traits) can engage with tasks that are properly representative of the decision-making scenario. The issue with non-realistic AI systems also limits the extraction of realistic behavior [145]. Many studies [5, 23, 94, 156] acknowledge this limitation, namely the low relevance of tasks to realistic scenarios.

4.3.5 AI System Fidelity. Studies examine human decision-making over a mix of AI models, from basic to advanced ML and deep learning ones (Table 7), depending on the task, dataset, and context. However, around 33% (19 studies) do not construct real AI interventions, and, instead, use **simulated AI or Wizard-of-Oz** [37] approaches, trying to replicate what users would believe or experience while interacting with AI-assisted decision-making. Although this is a well-accepted strategy to evaluate systems in HCI research, it risks not fully replicating a realistic AI system's behavior. Interventions also use hand-picked tasks to control the experimental

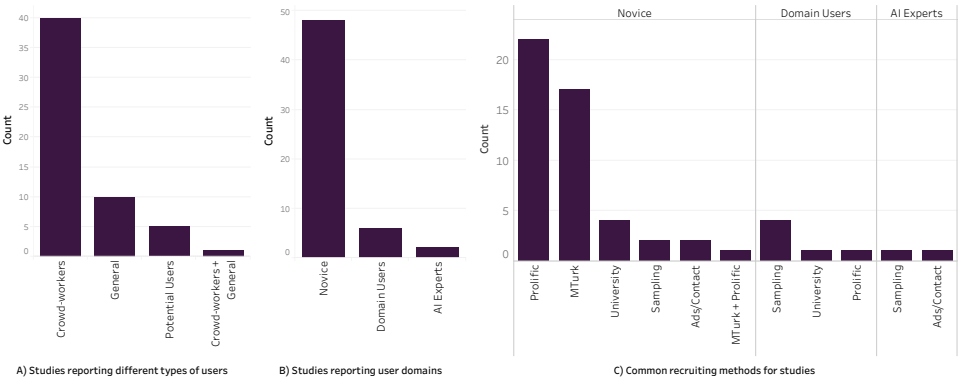

Figure 6: A) The majority of studies (40) are conducted with crowd workers and the general population, while there is a scarcity in research to recruit potential or actual users of the target application. B) Most studies (48) report that most users are novices, meaning they do not have any expertise to engage with tasks, while a few (6) use users with some domain experience, which is mostly in the healthcare domain. C) As evident from users' type, most studies recruit crowd workers from Prolific or Mechanical Turk, while a few recruit university students. Domain experts are mostly recruited through sampling methods.

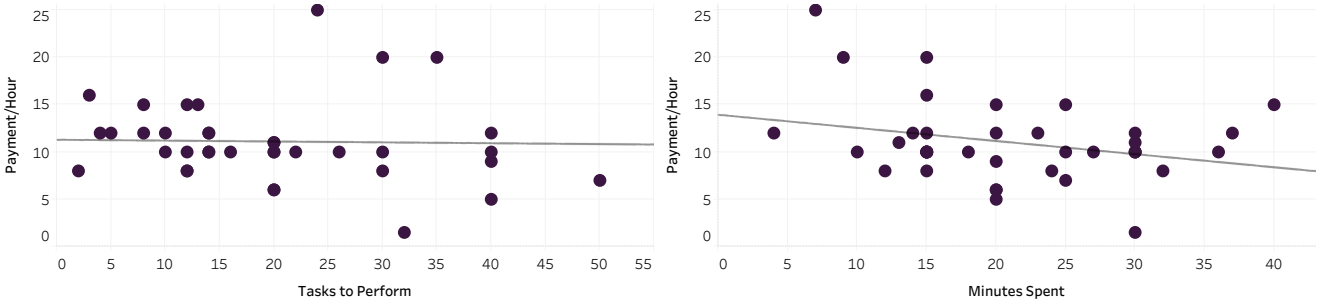

Figure 7: Left: Patterns of USD-adjusted hourly payment rate of crowd workers involved in human-AI reliance studies, compared to the number of tasks they are asked to perform (ranging from $1.5 to $25 and from 2 to 50 tasks in most studies, excluding two outliers of 100 and 256 tasks paid at $10/hour). We observe that the hourly payment remains relatively the same, regardless of the tasks the crowd workers are asked to complete. Right: Patterns of USD-adjusted hourly payment rate compared to the minutes spent during the experiment. We observe that the payment rate slightly decreases with increasing time spent. Crowd workers are expected to complete more tasks for a lower monetary reward. Overall, using lower monetary incentives for more tasks, as well as scenarios that are misaligned with the users' expertise, risks compromising human-AI reliance measurements.

conditions and replicate the specific performance, which limits researchers' understanding of human decision-making towards the real AI's *stochastic* behavior.

In fact, AI performance is often different when using simulated versus realistic or representative tasks, superficial versus real-world situations, and simple versus complex tasks [104, 127]. The literature acknowledges that evaluating novel AI advances through human subject experiments that involve realistic tasks is expensive in terms of both time and resources [9, 44], which may explain why the majority of relevant studies involve crowd workers and simplified tasks. However, **using fictitious or non-representative tasks**, such as predicting speed dating [168] or simulating judicial decisions [55], especially when performed with crowd workers in low-stakes or irrelevant contexts (about the task at hand), may not represent real-world situations.

Key insights. Studies have tested human reliance on AI in various application domains, including healthcare, business, and education. However, the majority of the studies employed **novice participants and fictive task scenarios**. Involving novice users, who – as opposed to experts – have limited expertise relevant to the task (e.g., diagnosis, recidivism, loan/income predictions), may considerably undermine the applicability of the results in real-world scenarios. The recruited crowd workers perform many tasks in the experiments within a short time, compared to non-crowd workers. Many studies acknowledge the limitations of involving crowd workers, and a few seek to enhance the recruited workers' expertise through domain knowledge or to increase their **attention** by adding incentives. However, incentives are not positively correlated with the number of tasks the crowd workers are expected to perform, which could perhaps be one of the reasons they fail to appropriately engage with the AI system at hand. Finally, many

Table 4: Common interventions and their impact on achieving appropriate reliance. Positive impact means that the respective methods were found useful in improving users' behavior for appropriate reliance (within specific experimental conditions). Likewise, the negative impact indicates that applying the respective method/concept decreased reliance under the experimental conditions.

Intervention and Study	Impact
Cognitive Forcing [10, 40, 71], Deductive-XAI [9, 27, 28], Second Opinion [32, 93], Natural Language XAI [76, 103], Interactive Tutorials [33, 63], Cognitive Cost (Effort) [151]	Positive
Task Difficulty [23, 127, 151], Negative First Impression [61, 108, 153], Poor Self Assessment [63, 94], Lack of Domain Knowledge [103, 156], Feature-based XAI [30, 170]	Negative
Example-based XAI [10, 29, 131, 156], Feature-based XAI [29, 133, 164], Showing AI Uncertainty [24, 98]	No Effect

studies do not employ real AI systems; rather, they rely on **simulated systems or Wizard-of-Oz** approaches, which further limit their operationalization prospect in real-world settings.

4.4 Prospective Themes in Human-AI Reliance

Here, we provide a bird's-eye view of the prospective themes in human-AI reliance, as well as an overview of approaches mentioned in the literature. Table 4 summarizes the most commonly applied interventions, the relevant studies, and their impact on achieving appropriate reliance.

Many studies [5, 10, 102, 156, 157, 170] report that **explanations increase users' over-reliance** on AI advice. Cognitive forcing [10, 40] and frictional methods [13, 15] are reported to reduce over-reliance. Vasconcelos et al. [151] further show that **explanations can reduce over-reliance without cognitive forcing**. For instance, if a task is challenging and its explanation is simple, then over-reliance can be reduced [151]. People weigh the potential benefits of cognitive effort against its perceived cost [78], a phenomenon also referred to as *cognitive-effort-discounting* [160]. For example, people are less likely to over-rely on the AI system's advice when the benefit of obtaining the correct answer substantially outweighs the risk of over-reliance (e.g., avoiding sending an unprofessional email to a boss), compared with situations in which the benefit-risk tradeoff is low (e.g., assessing the sentiment of a tweet) [151]. By reducing the cost of verifiability and plausibility of XAI techniques, decision-makers gain a better understanding of AI advice, potentially leading to better AI reliance [127].

Similar to cognitive forcing, partial explanations [40] can also help reduce over-reliance. Partial explanations align with a larger body of research on deliberation that encourages users to reflect on their decisions [40, 96]. Lu et al. [93] pointed out that **active solicitation** can also help reduce over-reliance, for instance, by getting a second opinion to form a consensus. Cabitza et al. [17] explored providing alternative advice (examples), which engages users to identify correct examples. Jeon et al. [71] explored the principles of logical validity, such as establishing a clear cause-and-effect relationship. He et al. [60] used an XAI dashboard, providing interactive, on-demand, and conversational explanations, thereby improving the users' understanding of the model [60]. Swaroop et al. [145] posit adapting AI advice to the over-reliance rate. Different people trust AI differently [145], and their traits and real-time behavior can help extract patterns that can inform the design to curb over-reliance and under-reliance. Sequential decisions where decision-makers build their perception of the AI with task progress have also been useful [123].

Task difficulty is not typically considered in studies [126]; however, it significantly impacts user reliance on AI suggestions. Task complexity increases information overload pertaining to the user's perception and capabilities [113]. Delegating easy tasks to AI alone can also reduce the burden on users to decide every task [144, 145]. He et al. [61] use debugging as a training method to help users learn tasks and AI mechanics. However, when users lack expertise in a task, they tend to (inappropriately) rely heavily on AI advice, especially as the task becomes more complex [38, 127]. Research suggests that when the task is difficult, the design should slow down users to help them make better decisions [72, 135].

Bansal et al. [5] posit that having the **agency** to correct AI advice can improve user performance. The users' accurate self-assessment of their skills/knowledge is also important. Buçinca et al. [11] highlighted that AI should reinforce **skill development and competence** in decision-makers, as current XAI often fails to address the knowledge gaps that users seek to fill [11]. They [11] showed that enhancing users' skills is important for improving engagement and human decision-making. Studies also showed that confirming decisions by expert or empirical validation can be useful, in the form of conceptual validations (CVs) [112]. User demographics (e.g., education, literacy) [33, 63, 168], and more importantly, domain expertise [43] also showed some effects on building reliance. Overall, studies show mixed results towards improving reliance on AI, and cite limitations in forming **appropriate human-AI reliance.**

Key insights. Certain methods, such as cognitive forcing, deductive XAI, and second opinions, show positive effects in building appropriate reliance. However, studies also indicate that several other applied interventions have either a negative or, at best, a neutral effect on appropriate reliance. In particular, the users' poor self-assessment (i.e., an overestimation of their skills) and the difficulty of the task can negatively impact their reliance on AI. These results can be attributed to several factors, such as mismatched user expertise as well as task relevance.

5 Discussion

Current human-AI decision-making research focuses on two main directions: 1) to engage users in evaluating AI more thoroughly, by

introducing frictions, and 2) to ease users' interaction, by simplifying their decisions. Several studies aim to enhance user engagement by implementing interventions that encourage them to use *analytical thinking* [10, 13, 40, 103] instead of blindly accepting AI advice. **Decision-making requires thoughtful engagement;** otherwise, users make flawed decisions or do not exhibit realistic behavior. Studies try to replicate human behavior with realistic applications; however, users often lack sufficient domain knowledge to engage with tasks. Many studies focus on identifying unified frameworks while acknowledging the limitations of crowd workers. Research explores how users behave under certain conditions, such as by investigating their *biases* [107, 123] or (self) *capabilities* [11, 17]. Researchers also explore *calibrating* the users' capabilities to curb under-reliance [24, 71, 97, 98] by making users aware of their biases. However, current HCI research shows that inappropriate reliance on AI advice is widespread. The field is yet to settle on rigorous metrics [13, 103, 132]. Research informs ways that the HCI community can contribute to a better understanding of appropriate reliance, particularly regarding the design of AI systems in real-world settings. In what follows, we reflect on our analysis and discuss the challenges explored in current human-AI reliance research.

5.1 User Engagement and Frictional Design

HCI researchers draw on methods from psychology (e.g., *dual-system theory* [72, 159], *reflective design* [36, 101, 109]) to capture users' attention and ensure their analytical engagement, and this way reduce over-reliance on AI advice, as evidenced by cognitive forcing and frictional AI designs [10, 13, 15, 40]. In general, a certain degree of **over-reliance is unavoidable by default** as humans tend to leverage system 1 (fast) thinking more than system 2 (deliberative) thinking, for the sake of *efficiency*. Hence, studies [10, 13, 15, 40] recommend *design interventions* to enhance oversight and discourage automatic behavior, such as through responsible nudging [25, 105, 119]. While nudging and frictions can induce behavior change, research emphasizes that they must be applied carefully to avoid manipulating user behavior through dark patterns or reducing disconnect from AI advice (e.g., due to algorithmic aversion) [16, 51]. Hence, it is critical to ensure that frictions or nudges are *transparent*, and that they indeed enhance task accuracy/performance rather than just targeting user alignment (agreement) with the system. Applying *"positive friction"* [31] carefully can also help **discourage the default human tendency for over-reliance**, while adhering to ethical values. Nevertheless, the exact way that people engage with AI may still depend on *situational factors* (e.g., time pressure) [144], *traits* (e.g., expertise) [43, 120], or *individuality* (e.g., motivation [10]).

Although decision-makers can be motivated or nudged to engage with AI, how they choose to rely on the AI system remains largely an unanswered question [30]. Our review shows that typical *AI system design* tries to favor users, making it **easier for them to align with AI advice than contesting it** (e.g., AI predicting X, providing the evidence while burdening users to find counter-evidence). Cabitza et al. [13] studied the influence AI exerts on users to align with its decision. Positive AI exertion, though, can help non-experts achieve higher performance with AI. Still, whether *technology dominance* has a positive effect or not is determined by how the interaction is designed. For instance, systems that provide full-fledged answers might induce over-dependence. Novices (often crowd workers) can be more prone to over-reliance and technology dominance, as they engage less analytically with it [30, 143].

Our review identifies that interaction should deter the users' *loss of skills* (over-dependence) and blind faith in AI due to its alleged *exertion* [15, 95]. Designing experiences that intentionally introduce frictions can **help users think slowly and deliberatively**. However, frictions can also lead to lower engagement because users will resist using systems with frictions; hence, design adaptations are necessary to apply them effectively [106]. For instance, engaging users to present multiple and contrastive examples instead of AI's advice on the decision implements cognitive friction while reducing disengagement [11, 14]. Reflective design [35], agency for deliberation [96], and evaluative AI [101] approaches can also be applied to prompt users to reflect more critically over their decision-making strategies. **"Reflective design"** is a highly interesting research avenue, which remains underexplored in our corpus.

5.2 Focus on User Factors or Biases

Studies in our corpus show that various biases could negatively influence the users' decisions, undermining their self-efficacy and AI perceptions [18, 39, 115]. Methods to reduce the effect of biases are also proposed to evaluate users' *confidence* in their self-assessments and expertise. Evaluating one's performance is important as **people tend to estimate their abilities poorly** [46, 70]. In addition, HCI researchers derive other human factors, such as *motivation*, *personality*, and economic *incentivization*, to adapt AI advice. Adapting AI based on user factors improves how AI advice is consumed by end users. One prominent example is identifying the users' personality [144, 145] and reliance strategy to subsequently adapt AI assistance for over-reliant users (i.e., by limiting AI advice).

We found that effective human-AI interaction depends on the users forming a faithful *mental representation* of AI and assessing the strengths and limitations of models. In some cases, studies highlight that effective interaction is essential to enhance the users' capabilities to understand data (e.g., to do practice tests to enhance their literacy). Hence, calibrating users' self-confidence is important to accommodate decisions with varying levels of uncertainty while reflecting on AI's limitations. Calibrating can be *appropriate* when the design supports users to rely on AI when it has higher confidence, and rely on themselves when it has low confidence in the decisions [97, 98].

Inappropriate self-assessment is only one of many causes of inappropriate reliance [98, 112]. Our review identifies a mismatch between real-world tasks and user knowledge/expertise, which may not mirror natural decision-making scenarios. Crowd workers can be involved for simple tasks [63], which most laypeople are capable of dealing with; however, actual applications have more factors, such as user expertise, familiarity, and stake. XAI methods also influence behavior, for instance, by showing false *depth of reasoning* to make users believe they understand models, which also affects users' perceived expertise under uncertainty. Still, designing XAI to reduce biases and lack of expertise continues to be a challenge in human-AI reliance research studies [80, 104, 107].

5.3 Expertise and Task Complexity

Our findings show that studies use simpler tasks and involve non-representative users due to their ease of availability on crowdsourcing platforms. However, it is also noted that crowd workers interact with tasks with very low *motivation* and with low *deliberation* [28, 71, 98]. In addition, decision-making with simple tasks does not necessarily generalize to high-stakes ones, and going beyond simple tasks requires domain expertise [112]. Our findings show significant differences between crowd workers' and domain experts' time on tasks, where crowd workers go through tasks more quickly. Even when the given task itself is high-stakes (e.g., recidivism [32]) or resembles real-world, if the users involved bear no responsibility or accountability or lack expertise [53, 152], the task may still fail to extract user reliance appropriately. *Marginal incentives* are not influential motivators to engage users with difficult tasks. Although reward can be a factor to motivate users, over-reliance is reduced when people see a value in completing the task. On the contrary, Grgić-Hlača et al. [55] argue that experiments with real users also lack realistic aspects, which is comparable to laypeople, i.e., failing to capture attention or expertise.

Many studies in our corpus used simpler tasks that lack real-world applicability and were generally designed to elicit user behavior without requiring domain expertise. Comparing the differences between the expertise of participants, Chen et al. [30] show the process through which decision-makers determine whether to rely on AI or not, e.g., a radiologist looking at an X-ray knows what to look for, and may form an opinion about the diagnosis, where a novice might just agree with the AI advice. Though studies [43, 62, 94] show that users' expertise with tasks can be enhanced through knowledge and training, it can have inherent limitations in extracting realistic human behavior having no stake in the decisions. Therefore, research needs to explore the tasks that are at the proper level of difficulty, which are challenging enough and require continued attention for the target users.

5.4 Understanding User Problems

Our review identifies that most current studies focus on applying different tasks and ML systems, and little attention has been paid so far to understanding the users' *actual* problems. Proxy tasks can lead users to analyze only the AI effect, while realistic tasks focus on improving decision-making, whether or not AI is used. For instance, Lee and Chew [88] show that AI experiments in the wild can elicit realistic user behavior better than laboratory-based ones. Using AI uncertainty and individual differences, it is important to identify when AI advice is needed and when users can rely on themselves to make decisions. When users are *capable* of executing the tasks themselves, they engage with AI advice *appropriately*, and interactions (back and forth) can support understanding uncertainties in AI advice. Users are better at calibrating their reliance on AI by *reasoning* or building *causality* [116] rather than using statistical or associative explanations. In addition, formulating a good *mental model* can facilitate enhancing user interaction with AI [66].

High focus on simplified tasks and simulated AI does not necessarily test interventions beyond crowd workers. Application-grounded experiments need to build realistic interactive AI systems so that users can appropriately rely on and override them when they are likely to fail. We see that engaging users with AI assistance can help to enhance their stakes in the decisions, as explored in earlier works [45, 121]. With step-wise examination of the task, users can build explanations and rationale with the system, which can help them contest, take ownership of their decisions, and establish critical engagement or mindful usage [64]. For instance, most existing work focuses on XAI; however, this is just one component of the ML pipeline, which also includes data, training, and evaluation. Understanding and supporting how people interact with these other components is therefore essential, and perhaps even more relevant than merely knowing how the model works [45, 121].

6 Recommendations for Further Studies

Current research on human-AI reliance has several open questions that researchers can further explore. In this section, we discuss the identified gaps and potential recommendations for future avenues. We provide practical pointers for designing interventions that can effectively elicit user behavior that is representative of real-world scenarios.

6.1 Emerging Research Gaps

Our analysis maps out **disparities in interventions** for both non-representative and realistic tasks. One overlooked issue is the task-expertise mismatch, i.e., a misalignment between the complexity of the tasks and the end users' expertise [68]. Hence, there is a need for appropriate *human-grounded* studies, carefully evaluating tasks with users' expertise, for instance, using realistic tasks with actual users [1, 44]. Although studies explore different tasks and domains, there is still a need to focus on the unique challenges posed by complex tasks. More **complex tasks tend to require greater cognitive effort** [160], making individuals more likely to (over) rely on AI systems for assistance. Future research should consider incorporating methodologies that take *task relevancy* into account for user expertise.

Establishing definitions and terminologies for appropriate reliance is integral. We discussed three views explored in research: *traditional* [151], *appropriateness* [132], and *dominance* [13]. Although these views focus on objective reliance on AI, there is a need to unify them as frameworks. With these defined views, studies use fragmented (often self-defined) concepts and measures to capture the reliance. This can also be a factor in missing representative studies covering human-AI reliance due to using non-uniform terms. Further, we recommend that the HCI community standardize the metrics for measuring user reliance. Future work should enhance the connection to established terminologies and measures, and explicitly build upon them to define newer adaptations [103].

To encourage users to maintain an active and engaged role in the decision-making while fostering their abilities, it is necessary to **add limited inconvenience or delay in the design**, to facilitate what in this study we refer to as *"slow human-AI interaction"*. For instance, Cabitza et al. [15] posit providing possible decisions instead of direct advice for user engagement, resonating with the psychological aspects of *reflective* [109] and *evaluative AI* [101]. Future research should focus on demonstrating its effectiveness in mitigating concerns of *automation bias* [108], *deskilling* [11], and

agency loss [121]. Our review finds that research should use a holistic perspective on algorithmic transparency, grounded in concrete *desiderata* [133] and encouraging thoughtful human-AI collaboration. Current human-AI reliance research does not fully represent the challenges and dynamics of real-world scenarios, particularly in complex tasks coupled with uncertainty.

Recent explorations show emerging directions; for instance, pointing out that people with different over-reliance qualities can be given different AI assistance [145]. A fixed type of AI assistance for all people and tasks can worsen decision accuracy compared to the accuracy achieved when relying solely on AI in many settings [12, 54], often because humans are over-relying on the AI. Quantifying the end-users' mental models of AI systems can help design effective decision aids, which can – in turn – empower and nudge people to rely on AI appropriately. Recently explored conversational XAI shows potential; however, it is reported to *oversimplify* the explanations and create an *illusion* of explanatory depth [60]. Another emerging issue concerns *deceptive* explanations [103], where research illustrates differences for correct and incorrect XAI. Also, there is a growing interest in understanding the decision-maker's expertise [88, 164]; hence, research should also consider answering such problems.

6.2 Recommendations for Interventions

Here, we discuss recommendations for interventions that facilitate eliciting user behavior to study human-AI reliance.

6.2.1 Use Reflective and Layered Disclosure. We find that most studies overload users with information (tasks, XAI, uncertainty, time-pressure, etc.). For effective interaction, a gradual/layered approach can assist users to contemplate and engage effectively, such as showing only the high-level information first and letting the user drill down for detailed evidence. Cox et al. [36] propose adding reflection design points [135] to create friction that helps users shift from system 1 to system 2 thinking. Reflections can include partial explanations, progressive disclosure, layering information, or thought-provoking questioning [142]. For example, interventions could include putting barriers on default functionalities, making them difficult to run, or slowing them down [58] to enhance deliberate thinking. Design frictions [35] can be useful for forming reliance; however, they should be carefully applied to avoid disengaging the users or exerting technology dominance [51].

6.2.2 Engage More Than Advice. Users can be engaged more effectively when they are given the option to contest and adjust an AI prediction rather than accepting it as-is [42]. Hence, enabling users to modify system outputs can be beneficial, such as in active/interactive learning [45, 136]. However, exposing the AI models' internal mechanics may not be effective at the beginning of the interaction, as it can overwhelm people; instead, it should be deployed in a step-wise approach, permitting people to dig down interactively to contest. The cognitive load incurred as a result of such a step-wise engagement approach can be reduced by guiding users to prevent them from reverting to System 1 thinking. Engaging also includes mapping people's stake in the task, for instance, co-designing decision support with relevant stakeholders and realistic (deployed) contexts. Interventions should aim to match the users' decisions with their mental models while ensuring that ethical concerns are addressed [118]. In general, transparent models are desirable due to ethical and trust issues.

6.2.3 Convey Limitations and Risks. Studies rarely [32, 43, 118] convey limitations or risks associated with AI advice. It is important to develop features that effectively reflect AI limitations or uncertainty. In practice, Prabhudesai et al. [118] show that suppressing uncertainty oversimplifies the tasks and can lead to over-reliance through what is referred to as the *"white-box paradox"* [13]; a situation where explanations increase automation bias, since users assume that XAI provides them with complete system understanding when in reality it does not. Intervention design should include detailed limitations and risks, for example, by including qualitative narratives, and uncertainty quantification of complex statistical decisions in a digestible (visual) manner [75]. Communicating uncertainty also helps users slow down and think analytically before making decisions. Conveying variations/uncertainty in the output, even when AI is wrong, can help users assess its limitations [131]. AI interventions should also enhance practical relevance to effectively impact human decision-making.

6.2.4 Support Reasoning and Interactive Hedging. It is essential to design methods that support reasoning and interactive hedging [8] for verifying multiple alternatives before trusting the outputs. For example, by providing evidence both for and against a potential decision, as opposed to providing a single decision recommendation, *judicial AI* [15] can reduce automation bias and support reasoning. Design should engage users to assess advice with evidence or alternatives, motivating the role of advisor and judge (JAS), and conduct further analysis instead of making instant decisions. These approaches can be further classified into cautious, analogical, or decentralized protocols, as instances of frictional AI [17]. This approach also requires enhancing the users' expertise in understanding and evaluating AI advice. Hence, it is important to consider to what extent users are aware of their abilities and those of the AI system.

6.2.5 Complement User Skills. Assessing users' expertise about the task is essential for eliciting a realistic behavior, for instance, to calibrate users' self-confidence in decision-making [97]. It is useful to design literacy interventions that enable users to analyze data distributions, outliers, or model parameters. Such literacy (becoming an expert through interactions) can help users rely appropriately [132]. Calibrating reliance and evaluating the performance of systems correctly is an essential and non-trivial task for users, which can enhance their comprehension of AI predictions and uncertainty. Interventions should also carefully select the explanatory reasoning style. This can include designing explanations with concrete objectives and providing cues [133]. Clear guidance on checking advice, knowing when to use it, and when to intervene manually can help users calibrate their reliance.

7 Limitations

This work has several limitations. First, like most reviews, we may have missed some relevant studies, even with a diverse search strategy and concrete inclusion criteria. Nevertheless, the quality of our corpus – comprising high-impact HCI venues such as PACMHCI,

CHI, IUI, and many similar venues – provides a comprehensive view of the relatively nascent topic of human-AI reliance. Human-AI decision-making studies can use highly variable measures and fluid measure definitions, which may have led to the exclusion of studies that place less focus on objective reliance. The corpus extraction also depended on search queries, which, although representative, may have missed articles that explore objective reliance but do not contain corresponding metadata. Our search was conducted using SCOPUS and the ACM Digital Library, similar to reviews in the related domain [81, 162]. We also used snowballing strategies to improve the wide coverage of studies. While our search and snowballing strategies were comprehensive, we acknowledge that including additional databases (such as IEEE Xplore) could further complement this study.

Second, our focus on the definition of objective reliance stems from Schemmer et al. [132], and the inclusion criteria were applied accordingly. Consequently, our approach for categorizing different methods of appropriate reliance (AI systems, users, interactions) represents only one of many possible categorization frameworks. Third, as we focus on HCI research, it could be possible that some work on human-AI reliance may exist outside the covered databases. Finally, despite careful evaluation, we acknowledge that results may be affected by screening bias and subjective interpretation during the application of inclusion and exclusion criteria.

8 Conclusion

Human-AI complementarity is still a far-fetched goal. Human reliance on AI advice is influenced by many factors, including the users' perception of AI systems, their cognitive factors, and how the AI advice is presented to them. HCI research has recently focused on the concept of appropriate human-AI reliance, differentiating it from user trust. In this work, we review 56 studies on human-AI reliance to examine how users can build appropriate reliance on AI advice, aiming to avoid both under- and over-reliance. Our findings show that there are clear gaps in the definitions used in the literature, as well as in measuring reliance in realistic contexts. Researchers use various experimental methods, often involving crowd workers, to study reliance, applying behavioral and interaction theories. Research shows progress towards identifying patterns for supporting reliance; however, there is a lack of common frameworks due to fragmented experiments across domains. We discuss several open-ended directions and emerging avenues for future research to better understand and design interventions that improve human reliance on AI advice. We provide our study corpus, coding, and analysis for future work to expand upon.

Acknowledgments

This work has been supported by Sappi, Geo Games, and the Media Research Lab at Rochester Institute of Technology.

References

[1] Abdulrahman Al-Surmi, Mahdi Bashiri, and Ioannis Koliousis. 2022. AI based decision making: combining strategies to improve operational performance. *International Journal of Production Research* 60, 14 (2022), 4464–4486. doi:10.1080/00207543.2021.1966540

[2] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion* 99 (2023), 101805. doi:10.1016/j.inffus.2023.101805

[3] OSF Anonymous. 2025. *Anonymized Material for Study: Review Protocol, Materials, Data Analysis, and Detailed Descriptions.* https://osf.io/sz6mf/?view_only=f143fac8f112423e96a02344c4a020be

[4] Annette Baier. 1986. Trust and antitrust. *ethics* 96, 2 (1986), 231–260.

[5] 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. doi:10.1145/3411764.3445717

[6] Astrid Bertrand, Rafik Belloum, James R. Eagan, and Winston Maxwell. 2022. How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) *(AIES '22)*. Association for Computing Machinery, New York, NY, USA, 78–91. doi:10.1145/3514094.3534164

[7] Nattapat Boonprakong, Benjamin Tag, Jorge Goncalves, and Tilman Dingler. 2025. How Do HCI Researchers Study Cognitive Biases? A Scoping Review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 473, 20 pages. doi:10.1145/3706598.3713450

[8] Michelle Brachman, Zahra Ashktorab, Michael Desmond, Evelyn Duesterwald, Casey Dugan, Narendra Nath Joshi, Qian Pan, and Aabhas Sharma. 2022. Reliance and Automation for Human-AI Collaborative Data Labeling Conflict Resolution. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 321 (Nov. 2022), 27 pages. doi:10.1145/3555212

[9] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 454–464. doi:10.1145/3377325.3377498

[10] Zana Buçinca, 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. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (April 2021), 21 pages. doi:10.1145/3449287

[11] Zana Buçinca, Siddharth Swaroop, Amanda E. Paluch, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2025. Contrastive Explanations That Anticipate Human Misconceptions Can Improve Human Decision-Making Skills. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1024, 25 pages. doi:10.1145/3706598.3713229

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

[13] Federico Cabitza, Andrea Campagner, Riccardo Angius, Chiara Natali, and Carlo Reverberi. 2023. AI Shall Have No Dominion: on How to Measure Technology Dominance in AI-supported Human decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 354, 20 pages. doi:10.1145/3544548.3581095

[14] Federico Cabitza, Andrea Campagner, Luca Ronzio, Matteo Cameli, Giulia Elena Mandoli, Maria Concetta Pastore, Luca Maria Sconfienza, Duarte Folgado, Marília Barandas, and Hugo Gamboa. 2023. Rams, hounds and white boxes: Investigating human–AI collaboration protocols in medical diagnosis. *Artificial Intelligence in Medicine* 138 (2023), 102506. doi:10.1016/j.artmed.2023.102506

[15] Federico Cabitza, Lorenzo Famiglini, Caterina Fregosi, Samuele Pe, Enea Parimbelli, Giovanni Andrea La Maida, and Enrico Gallazzi. 2025. From Oracular to Judicial: Enhancing Clinical Decision Making through Contrasting Explanations and a Novel Interaction Protocol. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 745–754. doi:10.1145/3708359.3712157

[16] Federico Cabitza, Caterina Fregosi, Andrea Campagner, and Chiara Natali. 2024. Explanations considered harmful: the impact of misleading explanations on accuracy in hybrid human-ai decision making. In *World conference on explainable artificial intelligence*. Springer, 255–269.

[17] Federico Cabitza, Chiara Natali, Lorenzo Famiglini, Andrea Campagner, Valerio Caccavella, and Enrico Gallazzi. 2024. Never tell me the odds: Investigating pro-hoc explanations in medical decision making. *Artificial Intelligence in Medicine* 150 (2024), 102819. doi:10.1016/j.artmed.2024.102819

[18] Ángel Alexander Cabrera, Adam Perer, and Jason I. Hong. 2023. Improving Human-AI Collaboration With Descriptions of AI Behavior. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 136 (April 2023), 21 pages. doi:10.1145/3579612

[19] John T Cacioppo and Richard E Petty. 1982. The need for cognition. *Journal of personality and social psychology* 42, 1 (1982), 116. doi:10.1037/0022-3514.42.1.116

[20] 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. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 104 (Nov. 2019), 24 pages. doi:10.1145/3359206

[21] Longbing Cao. 2022. AI in Finance: Challenges, Techniques, and Opportunities. *ACM Comput. Surv.* 55, 3, Article 64 (Feb. 2022), 38 pages. doi:10.1145/3502289

[22] Shiye Cao, Catalina Gomez, and Chien-Ming Huang. 2023. How Time Pressure in Different Phases of Decision-Making Influences Human-AI Collaboration. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 277 (Oct. 2023), 26 pages. doi:10.1145/3610068

[23] Shiye Cao and Chien-Ming Huang. 2022. Understanding User Reliance on AI in Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 471 (Nov. 2022), 23 pages. doi:10.1145/3555572

[24] Shiye Cao, Anqi Liu, and Chien-Ming Huang. 2024. Designing for Appropriate Reliance: The Roles of AI Uncertainty Presentation, Initial User Decision, and User Demographics in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 41 (April 2024), 32 pages. doi:10.1145/3637318

[25] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15. doi:10.1145/3290605.3300733

[26] Emma R. Casolin and Flora D. Salim. 2024. Towards Understanding Human-AI Reliance Patterns Through Explanation Styles. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Melbourne VIC, Australia) *(UbiComp '24)*. Association for Computing Machinery, New York, NY, USA, 861–865. doi:10.1145/3675094.3678996

[27] Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. 2023. Effects of AI and Logic-Style Explanations on Users' Decisions Under Different Levels of Uncertainty. *ACM Trans. Interact. Intell. Syst.* 13, 4, Article 22 (Dec. 2023), 42 pages. doi:10.1145/3588320

[28] Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. 2023. Supporting High-Uncertainty Decisions through AI and Logic-Style Explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 251–263. doi:10.1145/3581641.3584080

[29] Federico Maria Cau and Lucio Davide Spano. 2025. The Influence of Curiosity Traits and On-Demand Explanations in AI-Assisted Decision-Making. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 1440–1457. doi:10.1145/3708359.3712165

[30] Valerie Chen, Q. Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations. 7, CSCW2, Article 370 (Oct. 2023), 32 pages. doi:10.1145/3610219

[31] Zeya Chen and Ruth Schmidt. 2024. Exploring a behavioral model of "positive friction" in human-AI interaction. In *International Conference on Human-Computer Interaction*. Springer, 3–22. doi:10.1007/978-3-031-61353-1_1

[32] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2023. Are Two Heads Better Than One in AI-Assisted Decision Making? Comparing the Behavior and Performance of Groups and Individuals in Human-AI Collaborative Recidivism Risk Assessment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 348, 18 pages. doi:10.1145/3544548.3581015

[33] Chun-Wei Chiang and Ming Yin. 2022. Exploring the Effects of Machine Learning Literacy Interventions on Laypeople's Reliance on Machine Learning Models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 148–161. doi:10.1145/3490099.3511121

[34] Hyesun Choung, Prabu David, and Arun Ross. 2023. Trust in AI and Its Role in the Acceptance of AI Technologies. *International Journal of Human–Computer Interaction* 39, 9 (2023), 1727–1739. doi:10.1080/10447318.2022.2050543

[35] NAJ Cornelissen, RJM Van Eerdt, HK Schraffenberger, and Willem FG Haselager. 2022. Reflection machines: increasing meaningful human control over Decision Support Systems. *Ethics and Information Technology* 24, 2 (2022), 19. doi:10.1007/s10676-022-09645-y

[36] Anna L Cox, Sandy JJ Gould, Marta E Cecchinato, Ioanna Iacovides, and Ian Renfree. 2016. Design frictions for mindful interactions: The case for microboundaries. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*. 1389–1397. doi:10.1145/2851581.2892410

[37] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies: why and how. In *Proceedings of the 1st international conference on Intelligent user interfaces*. Association for Computing Machinery, 193–200. doi:10.1145/169891.169968

[38] Devleena Das and Sonia Chernova. 2020. Leveraging rationales to improve human task performance. In *Proceedings of the 25th international conference on intelligent user interfaces*. 510–518.

[39] Tushar Kanti Das and Bing-Sheng Teng. 1999. Cognitive biases and strategic decision processes: An integrative perspective. *Journal of management studies* 36, 6 (1999), 757–778.

[40] Sander de Jong, Ville Paananen, Benjamin Tag, and Niels van Berkel. 2025. Cognitive Forcing for Better Decision-Making: Reducing Overreliance on AI Systems Through Partial Explanations. 9, 2, Article CSCW048 (May 2025), 30 pages. doi:10.1145/3710946

[41] Dominik Dellermann, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister. 2019. Hybrid intelligence. *Business & Information Systems Engineering* 61, 5 (2019), 637–643. doi:10.1007/s12599-019-00595-2

[42] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General* 144, 1 (2015), 114.

[43] Murat Dikmen and Catherine Burns. 2022. The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies* 162 (2022), 102792. doi:10.1016/j.ijhcs.2022.102792

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

[45] John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning. *ACM Trans. Interact. Intell. Syst.* 8, 2, Article 8 (June 2018), 37 pages. doi:10.1145/3185517

[46] David Dunning, Chip Heath, and Jerry M Suls. 2004. Flawed self-assessment: Implications for health, education, and the workplace. *Psychological science in the public interest* 5, 3 (2004), 69–106.

[47] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies* 58, 6 (2003), 697–718. doi:10.1016/S1071-5819(03)00038-7 Trust and Technology.

[48] Jessica Maria Echterhoff, Matin Yarmand, and Julian McAuley. 2022. AI-Moderated Decision-Making: Capturing and Balancing Anchoring Bias in Sequential Decision Tasks. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 161, 9 pages. doi:10.1145/3491102.3517443

[49] Sven Eckhardt, Niklas Kühl, Mateusz Dolata, and Gerhard Schwabe. 2025. A survey of AI reliance. *Comput. Surveys* 58, 6 (2025), 1–37.

[50] 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. doi:10.1007/978-3-030-60117-1_33

[51] Upol Ehsan and Mark O Riedl. 2024. Explainability pitfalls: Beyond dark patterns in explainable AI. *Patterns* 5, 6 (2024).

[52] Christoph Engel, Andreas Glöckner, and Sinika Timme. 2017. *Defendant should have the last word: Experimentally manipulating order and provisional assessment of the facts in criminal procedure.* Technical Report. Preprints of the Max Planck Institute for Research on Collective Goods.

[53] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine* 4, 1 (2021), 31.

[54] Ben Green and Yiling Chen. 2019. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 50 (Nov. 2019), 24 pages. doi:10.1145/3359152

[55] Nina Grgić-Hlača, Christoph Engel, and Krishna P. Gummadi. 2019. Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 178 (Nov. 2019), 25 pages. doi:10.1145/3359280

[56] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* 51, 5, Article 93 (Aug. 2018), 42 pages. doi:10.1145/3236009

[57] Ziyang Guo, Yifan Wu, Jason D. Hartline, and Jessica Hullman. 2024. A Decision Theoretic Framework for Measuring AI Reliance. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 221–236. doi:10.1145/3630106.3658901

[58] Lars Hallnäs and Johan Redström. 2001. Slow technology–designing for reflection. *Personal and ubiquitous computing* 5, 3 (2001), 201–212.

[59] PA Hancock, Theresa T Kessler, Alexandra D Kaplan, Kimberly Stowers, J Christopher Brill, Deborah R Billings, Kristin E Schaefer, and James L Szalma. 2023. How and why humans trust: A meta-analysis and elaborated model. *Frontiers in psychology* 14 (2023), 1081086. doi:10.3389/fpsyg.2023.1081086

[60] Gaole He, Nilay Aishwarya, and Ujwal Gadiraju. 2025. Is Conversational XAI All You Need? Human-AI Decision Making With a Conversational XAI Assistant. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 907–924. doi:10.1145/3708359.3712133

[61] Gaole He, Abri Bharos, and Ujwal Gadiraju. 2024. To Err Is AI! Debugging as an Intervention to Facilitate Appropriate Reliance on AI Systems. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media* (Poznan, Poland) *(HT '24)*. Association for Computing Machinery, New York, NY, USA, 98–105. doi:10.1145/3648188.3675130

[62] Gaole He, Stefan Buijsman, and Ujwal Gadiraju. 2023. How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 276 (Oct. 2023), 29 pages. doi:10.1145/3610067

[63] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 113, 18 pages. doi:10.1145/3544548.3581025

[64] Maren Hinrichs, Thi Bich Diep Bui, and Stefan Schneegass. 2024. Exploring the Effects of User Input and Decision Criteria Control on Trust in a Decision Support Tool for Spare Parts Inventory Management. In *Proceedings of the International Conference on Mobile and Ubiquitous Multimedia*. 313–323.

[65] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors* 57, 3 (2015), 407–434. doi:10.1177/0018720814547570 PMID: 25875432.

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

[67] Robin M Hogarth and Emre Soyer. 2015. Providing information for decision making: Contrasting description and simulation. *Journal of Applied Research in Memory and Cognition* 4, 3 (2015), 221–228.

[68] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 2019. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300809

[69] Kori Inkpen, Shreya Chappidi, Keri Mallari, Besmira Nushi, Divya Ramesh, Pietro Michelucci, Vani Mandava, Libuše Hannah Vepřek, and Gabrielle Quinn. 2023. Advancing Human-AI Complementarity: The Impact of User Expertise and Algorithmic Tuning on Joint Decision Making. *ACM Transactions on Computer-Human Interaction* 30, 5 (2023), 1–29.

[70] Rachel A Jansen, Anna N Rafferty, and Thomas L Griffiths. 2021. A rational model of the Dunning–Kruger effect supports insensitivity to evidence in low performers. *Nature Human Behaviour* 5, 6 (2021), 756–763.

[71] Youngseung Jeon, Christopher Hwang, and Xiang 'Anthony' Chen. 2025. Empowering Medical Data Labeling for Non-Experts with DANNY: Enhancing Accuracy and Mitigating Over-Reliance on AI. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 624–640. doi:10.1145/3708359.3712161

[72] Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.

[73] Daniel Kahneman and Gary Klein. 2009. Conditions for intuitive expertise: a failure to disagree. *American psychologist* 64, 6 (2009), 515.

[74] Patricia K. Kahr, Gerrit Rooks, Martijn C. Willemsen, and Chris C. P. Snijders. 2024. Understanding Trust and Reliance Development in AI Advice: Assessing Model Accuracy, Model Explanations, and Experiences from Previous Interactions. *ACM Trans. Interact. Intell. Syst.* 14, 4, Article 29 (Dec. 2024), 30 pages. doi:10.1145/3686164

[75] 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* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376219

[76] Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 822–835. doi:10.1145/3630106.3658941

[77] 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3290605.3300641

[78] Wouter Kool and Matthew Botvinick. 2018. Mental labour. *Nature human behaviour* 2, 12 (2018), 899–908.

[79] Justin Kruger and David Dunning. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology* 77, 6 (1999), 1121.

[80] 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). doi:10.48550/ARXIV.1902.00006

[81] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 2023. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1369–1385. doi:10.1145/3593013.3594087

[82] 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* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376873

[83] 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* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 29–38. doi:10.1145/3287560.3287590

[84] J. Richard Landis and Gary G. Koch. 1977. An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers. *Biometrics* 33, 2 (1977), 363–374. http://www.jstor.org/stable/2529786

[85] Nancy K Lankton, D Harrison McKnight, and John Tripp. 2015. Technology, humanness, and trust: Rethinking trust in technology. *Journal of the association for information systems* 16, 10 (2015), 1. doi:10.17705/1jais.00411

[86] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How we analyzed the COMPAS recidivism algorithm. ProPublica (2016). *URL: https://www. propublica. org/article/how-we-analyzed-the-compas-recidivism-algorithm* (2016).

[87] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80. doi:10.1518/hfes.46.1.50_30392 PMID: 15151155.

[88] Min Hun Lee and Chong Jun Chew. 2023. Understanding the Effect of Counterfactual Explanations on Trust and Reliance on AI for Human-AI Collaborative Clinical Decision Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 369 (Oct. 2023), 22 pages. doi:10.1145/3610218

[89] Benedikt Leichtmann, Christina Humer, Andreas Hinterreiter, Marc Streit, and Martina Mara. 2023. Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior* 139 (2023), 107539.

[90] Qing Li, Sharon Chu, Nanjie Rao, and Mahsan Nourani. 2020. Understanding the effects of explanation types and user motivations on recommender system use. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 83–91.

[91] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103. doi:10.1016/j.obhdp.2018.12.005

[92] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, Richard Jiang, Hassan Khosravi, Freddy Lecue, Gianclaudio Malgieri, Andrés Páez, Wojciech Samek, Johannes Schneider, Timo Speith, and Simone Stumpf. 2024. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion* 106 (2024), 102301. doi:10.1016/j.inffus.2024.102301

[93] Zhuoran Lu, Dakuo Wang, and Ming Yin. 2024. Does More Advice Help? The Effects of Second Opinions in AI-Assisted Decision Making. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 217 (April 2024), 31 pages. doi:10.1145/3653708

[94] 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 78, 16 pages. doi:10.1145/3411764.3445562

[95] Caitlin Lustig, Katie Pine, Bonnie Nardi, Lilly Irani, Min Kyung Lee, Dawn Nafus, and Christian Sandvig. 2016. Algorithmic authority: the ethics, politics, and economics of algorithms that interpret, decide, and manage. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 1057–1062.

[96] Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. 2025. Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 261, 23 pages. doi:10.1145/3706598.3713423

[97] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing

Machinery, New York, NY, USA, Article 759, 19 pages. doi:10.1145/3544548.3581058

[98] Shuai Ma, Xinru Wang, Ying Lei, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2024. "Are You Really Sure?" Understanding the Effects of Human Self-Confidence Calibration in AI-Assisted Decision Making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 840, 20 pages. doi:10.1145/3613904.3642671

[99] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. 2020. International evaluation of an AI system for breast cancer screening. *Nature* 577, 7788 (2020), 89–94.

[100] Siddharth Mehrotra, Chadha Degachi, Oleksandra Vereschak, Catholijn M. Jonker, and Myrthe L. Tielman. 2024. A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction: Trends, Opportunities and Challenges. *ACM J. Responsib. Comput.* 1, 4, Article 26 (Nov. 2024), 45 pages. doi:10.1145/3696449

[101] Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 333–342. doi:10.1145/3593013.3594001

[102] Katelyn Morrison, Donghoon Shin, Kenneth Holstein, and Adam Perer. 2023. Evaluating the Impact of Human Explanation Strategies on Human-AI Visual Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 48 (April 2023), 37 pages. doi:10.1145/3579481

[103] Katelyn Morrison, Philipp Spitzer, Violet Turri, Michelle Feng, Niklas Kühl, and Adam Perer. 2024. The Impact of Imperfect XAI on Human-AI Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 183 (April 2024), 39 pages. doi:10.1145/3641022

[104] Heleen Muijlwijk, Martijn C. Willemsen, Barry Smyth, and Wijnand A. IJsselsteijn. 2024. Benefits of Human-AI Interaction for Expert Users Interacting with Prediction Models: a Study on Marathon Running. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 245–258. doi:10.1145/3640543.3645205

[105] Mohammad Naiseh, Reem S Al-Mansoori, Dena Al-Thani, Nan Jiang, and Raian Ali. 2021. Nudging through friction: an approach for calibrating trust in explainable AI. In *2021 8th International Conference on Behavioral and Social Computing (BESC)*. IEEE, 1–5. doi:10.1109/BESC53957.2021.9635271

[106] Don Norman. 2013. *The design of everyday things: Revised and expanded edition*. Basic books.

[107] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 112–121.

[108] 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 *Proceedings of the 26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 340–350. doi:10.1145/3397481.3450639

[109] Jasminko Novak, Kalina Drenska, Ksenia Koroleva, Lukas Pfahler, Lavinia Marin, Judith Möller, Ozlem Özgöbek, Martijn C Willemsen, Dorothee Dersch, Enny Das, et al. 2021. Towards reflective AI: needs, challenges and directions for future research. (2021).

[110] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372 (2021). doi:10.1136/bmj.n71

[111] Raja Parasuraman and Dietrich H Manzey. 2010. Complacency and bias in human use of automation: An attentional integration. *Human factors* 52, 3 (2010), 381–410. doi:10.1177/0018720810376055

[112] Saumya Pareek, Niels van Berkel, Eduardo Velloso, and Jorge Goncalves. 2024. Effect of Explanation Conceptualisations on Trust in AI-assisted Credibility Assessment. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 383 (Nov. 2024), 31 pages. doi:10.1145/3686922

[113] Alison Parkes. 2017. The effect of individual and task characteristics on decision aid reliance. *Behaviour & Information Technology* 36, 2 (2017), 165–177.

[114] Lu Peng, Dailin Li, Zhaotong Zhang, Tingru Zhang, Anqi Huang, Shaohui Yang, and Yu Hu. 2024. Human-AI collaboration: Unraveling the effects of user proficiency and AI agent capability in intelligent decision support systems. *International Journal of Industrial Ergonomics* 103 (2024), 103629. doi:10.1016/j.ergon.2024.103629

[115] Gloria Phillips-Wren, Daniel J Power, and Manuel Mora. 2019. Cognitive bias, decision styles, and risk attitudes in decision making and DSS. 63–66 pages. doi:10.1080/12460125.2019.1646509

[116] Markus Plass, Michaela Kargl, Patrick Nitsche, Emilian Jungwirth, Andreas Holzinger, and Heimo Müller. 2022. Understanding and Explaining Diagnostic Paths: Toward Augmented Decision Making. *IEEE Computer Graphics and Applications* 42, 6 (2022), 47–57. doi:10.1109/MCG.2022.3197957

[117] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 237, 52 pages. doi:10.1145/3411764.3445315

[118] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q. Vera Liao, and Nikola Banovic. 2023. Understanding Uncertainty: How Lay Decision-makers Perceive and Interpret Uncertainty in Human-AI Decision Making. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 379–396. doi:10.1145/3581641.3584033

[119] Muhammad Raees, Vassilis-Javed Khan, and Konstantinos Papangelis. 2025. Exploring Persuasive Engagement to Reduce Over-Reliance on AI-Assistance in a Customer Classification Case. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*. doi:10.1145/3699682.3728320

[120] Muhammad Raees, Vassilis-Javed Khan, and Konstantinos Papangelis. 2025. Towards Understanding Persuasive and Personalized Engagement for Human-AI Reliance. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*. doi:10.1145/3708319.3733643

[121] Muhammad Raees, Inge Meijerink, Ioanna Lykourentzou, Vassilis-Javed Khan, and Konstantinos Papangelis. 2024. From explainable to interactive AI: A literature review on current trends in human-AI interaction. *International Journal of Human-Computer Studies* 189, September 2024 (2024), 103301. doi:10.1016/j.ijhcs.2024.103301

[122] Gonzalo Ramos, Christopher Meek, Patrice Simard, Jina Suh, and Soroush Ghorashi. 2020. Interactive machine teaching: a human-centered approach to building machine-learned models. *Human–Computer Interaction* 35, 5-6 (2020), 413–451.

[123] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. 2022. Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 83 (April 2022), 22 pages. doi:10.1145/3512930

[124] Amy Rechkemmer and Ming Yin. 2022. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 535, 14 pages. doi:10.1145/3491102.3501967

[125] Lee Ross and Andrew Ward. 2013. Naive realism in everyday life: Implications for social conflict and misunderstanding. In *Values and knowledge*. Psychology Press, 103–135.

[126] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2023. A Missing Piece in the Puzzle: Considering the Role of Task Complexity in Human-AI Decision Making. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization* (Limassol, Cyprus) *(UMAP '23)*. Association for Computing Machinery, New York, NY, USA, 215–227. doi:10.1145/3565472.3592959

[127] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2024. Dealing with Uncertainty: Understanding the Impact of Prognostic Versus Diagnostic Tasks on Trust and Reliance in Human-AI Decision Making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 25, 17 pages. doi:10.1145/3613904.3641905

[128] Suprateek Sarker, Sutirtha Chatterjee, Xiao Xiao, and Amany Elbanna. 2019. The Sociotechnical Axis of Cohesion for the Is Discipline: Its Historical Legacy and its Continued Relevance. *MIS Quarterly* 43, 3 (2019), pp. 695–720, A1–A5. https://www.jstor.org/stable/26848052

[129] 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* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 240–251. doi:10.1145/3301275.3302308

[130] Nicolas Scharowski, Sebastian AC Perrig, Melanie Svab, Klaus Opwis, and Florian Brühlmann. 2023. Exploring the effects of human-centered AI explanations on trust and reliance. *Frontiers in Computer Science* 5 (2023), 1151150. doi:10.3389/fcomp.2023.1151150

[131] Max Schemmer, Andrea Bartos, Philipp Spitzer, Patrick Hemmer, Niklas Kühl, Jonas Liebschner, and Gerhard Satzger. 2023. Towards effective human-AI decision-making: The role of human learning in appropriate reliance on AI advice. In *Forty-Fourth International Conference on Information Systems* (Hyderabad, India) *(ICIS 2023 Proceedings)*. AISEL, 1–17. https://aisel.aisnet.org/

icis2023/hti/hti/14
[132] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 410–422. doi:10.1145/3581641.3584066
[133] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl. 2024. Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 836, 18 pages. doi:10.1145/3613904.3642621
[134] Jakob Schoeffer, Johannes Jakubik, Michael Vössing, Niklas Kühl, and Gerhard Satzger. 2025. AI reliance and decision quality: Fundamentals, interdependence, and the effects of interventions. *Journal of Artificial Intelligence Research* 82 (2025), 471–501. doi:10.1613/jair.1.15873
[135] Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph'Jofish' Kaye. 2005. Reflective design. In *Proceedings of the 4th decennial conference on Critical computing: between sense and sensibility*. 49–58.
[136] Burr Settles. 2009. Active learning literature survey. (2009).
[137] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.
[138] Ben Shneiderman. 2020. Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction* 12, 3 (2020), 109–124. doi:10.17705/1thci.00131
[139] Keng Siau and Weiyu Wang. 2018. Building trust in artificial intelligence, machine learning, and robotics. *Cutter business technology journal* 31, 2 (2018), 47–53.
[140] Janet A Sniezek and Timothy Buckley. 1995. Cueing and cognitive conflict in judge-advisor decision making. *Organizational behavior and human decision processes* 62, 2 (1995), 159–174.
[141] Hannah Snyder. 2019. Literature review as a research methodology: An overview and guidelines. *Journal of Business Research* 104 (2019), 333–339. doi:10.1016/j.jbusres.2019.07.039
[142] 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* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 107–120. doi:10.1145/3301275.3302322
[143] Steve G Sutton, Vicky Arnold, and Matthew Holt. 2022. An extension of the theory of technology dominance: understanding the underlying nature, causes and effects. *Causes and effects* (2022).
[144] Siddharth Swaroop, Zana Buçinca, Krzysztof Z. Gajos, and Finale Doshi-Velez. 2024. Accuracy-Time Tradeoffs in AI-Assisted Decision Making under Time Pressure. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 138–154. doi:10.1145/3640543.3645206
[145] Siddharth Swaroop, Zana Buçinca, Krzysztof Z. Gajos, and Finale Doshi-Velez. 2025. Personalising AI Assistance Based on Overreliance Rate in AI-Assisted Decision Making. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 1107–1122. doi:10.1145/3708359.3712128
[146] Heliodoro Tejeda, Aakriti Kumar, Padhraic Smyth, and Mark Steyvers. 2022. AI-assisted decision-making: A cognitive modeling approach to infer latent reliance strategies. *Computational Brain & Behavior* 5, 4 (2022), 491–508. doi:10.1007/s42113-022-00157-y
[147] Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. 2022. Capable but amoral? Comparing AI and human expert collaboration in ethical decision making. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.
[148] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. 2021. Second Chance for a First Impression? Trust Development in Intelligent System Interaction. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (Utrecht, Netherlands) *(UMAP '21)*. Association for Computing Machinery, New York, NY, USA, 77–87. doi:10.1145/3450613.3456817
[149] Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *science* 211, 4481 (1981), 453–458.
[150] Amos Tversky and Daniel Kahneman. 1989. Rational choice and the framing of decisions. In *Multiple criteria decision making and risk analysis using microcomputers*. Springer, 81–126.
[151] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 129 (April 2023), 38 pages. doi:10.1145/3579605
[152] Tiffany C Veinot, Hannah Mitchell, and Jessica S Ancker. 2018. Good intentions are not enough: how informatics interventions can worsen inequality. *Journal of the American Medical Informatics Association* 25, 8 (2018), 1080–1088.
[153] Mor Vered, Tali Livni, Piers Douglas Lionel Howe, Tim Miller, and Liz Sonenberg. 2023. The effects of explanations on automation bias. *Artificial Intelligence* 322 (2023), 103952. doi:10.1016/j.artint.2023.103952
[154] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 327 (Oct. 2021), 39 pages. doi:10.1145/3476068
[155] Lu Wang, Greg A. Jamieson, and Justin G. Hollands. 2008. Improving Reliability Awareness to Support Appropriate Trust and Reliance on Individual Combat Identification Systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 52, 4 (2008), 292–296. doi:10.1177/154193120805200420
[156] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 318–328. doi:10.1145/3397481.3450650
[157] Xinru Wang and Ming Yin. 2022. Effects of Explanations in AI-Assisted Decision Making: Principles and Comparisons. *ACM Trans. Interact. Intell. Syst.* 12, 4, Article 27 (Nov. 2022), 36 pages. doi:10.1145/3519266
[158] Xinru Wang and Ming Yin. 2023. Watch out for updates: Understanding the effects of model explanation updates in ai-assisted decision making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.
[159] Peter C Wason and J St BT Evans. 1974. Dual processes in reasoning? *Cognition* 3, 2 (1974), 141–154.
[160] Andrew Westbrook, Daria Kester, and Todd S Braver. 2013. What is the subjective cost of cognitive effort? Load, trait, and aging effects revealed by economic preference. *PloS one* 8, 7 (2013), e68210.
[161] Earl L Wiener. 1981. Complacency: Is the term useful for air safety. In *Proceedings of the 26th corporate aviation safety seminar*, Vol. 117. 116–125.
[162] Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. 2023. Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 755, 16 pages. doi:10.1145/3544548.3581197
[163] Claes Wohlin. 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering* (London, England, United Kingdom) *(EASE '14)*. Association for Computing Machinery, New York, NY, USA, Article 38, 10 pages. doi:10.1145/2601248.2601268
[164] Oskar Wysocki, Jessica Katharine Davies, Markel Vigo, Anne Caroline Armstrong, Dónal Landers, Rebecca Lee, and André Freitas. 2023. Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. *Artificial Intelligence* 316 (2023), 103839. doi:10.1016/j.artint.2022.103839
[165] Wei Xu. 2019. Toward human-centered AI: a perspective from human-computer interaction. *Interactions* 26, 4 (June 2019), 42–46. doi:10.1145/3328485
[166] Wei Xu, Marvin J Dainoff, Liezhong Ge, and Zaifeng Gao. 2023. Transitioning to human interaction with AI systems: New challenges and opportunities for HCI professionals to enable human-centered AI. *International Journal of Human–Computer Interaction* 39, 3 (2023), 494–518. doi:10.1080/10447318.2022.2041900
[167] 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* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 189–201. doi:10.1145/3377325.3377480
[168] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300509
[169] Qiaoning Zhang, Matthew L Lee, and Scott Carter. 2022. You Complete Me: Human-AI Teams and Complementary Expertise. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 114, 28 pages. doi:10.1145/3491102.3517791
[170] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. 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* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. doi:10.1145/3351095.3372852

A Summaries of Review

Table 5: Objective and subjective measures used in studies. The objective reliance measures are used to study how users fare with decisions. Trust is a commonly used subjective measure. Other measures are used to understand users' interaction with systems.

Category	Reported Measures
Reliance	Agreement Fraction [13, 22, 23, 28, 55, 60–63, 76, 88, 94, 96–98, 102, 123, 127, 156, 168, 170], Switch Fraction [22, 24, 60–63, 94, 96–98, 102, 112, 114, 127, 168, 170], AI-effect on Accuracy [5, 10, 11, 13, 15, 40, 43, 114, 131, 164, 170], Accuracy [8, 9, 11, 17, 18, 23, 27–29, 32, 40, 43, 60–63, 71, 76, 83, 88, 93, 96, 98, 108, 112, 118, 123, 127, 132, 133, 144–146, 156], Over-reliance [10, 27, 29, 32, 93, 96, 98, 103, 133, 144, 145, 151, 167], Under-reliance [27, 32, 93, 96, 98, 103, 133, 167], Human Error [10], Weight of Advice [13, 33, 117, 146], Absolute Percentage Error [33], RAIR [13, 60–63, 103, 127, 131, 132, 145], RSR [13, 60–63, 103, 127, 131, 132], AI Safety [13], Deception of Reliance [103], Accuracy-wid [60, 62, 63, 98, 127]
Trust	Subjective Trust Rating [8–10, 18, 23, 43, 76, 83, 88, 94, 96, 97, 102, 151, 164, 167, 168, 170], Trust in Automation [22, 40, 60–63, 112, 127], Affinity for Technology [60–63, 112, 127]
Others	Understanding [9, 24, 32, 33, 43, 60, 76, 88, 94, 96, 112, 156, 164, 167], Confidence Rating [15, 17, 23, 24, 27, 29, 32, 40, 55, 76, 93, 94, 96–98, 112, 114, 117, 132, 167], Usefulness/helpfulness [5, 9, 15, 17, 18, 22, 33, 62, 63, 88, 96, 97, 108, 112, 144, 164], Mental Demand (and/or Task Load) [9, 10, 29, 61, 88, 96–98, 108], Appropriateness [9, 11, 24, 27, 88, 112], Enjoyment [11, 33, 145, 151], Preference [10, 18, 29, 88, 145, 167], Complexity [10, 17, 40, 43, 96–98, 144], Satisfaction [33, 40, 88, 96–98, 164], Comfort [167], Assertiveness [103], Accountability [32], Agreement [23], Literacy [43], User Engagement [60], Risk [43], Fairness [133], Personality Traits [144, 145], Motivation [11, 145]

Table 6: Research uses various types of decision-making tasks in several real-world and toy domains, ranging from high-stakes (healthcare) to low-stakes (leisure), to study human reliance patterns.

Domains	Decision-Task
Medical/ Healthcare	Medical Diagnostic [13, 15, 17, 24, 71, 76, 153], Patient Seriousness Classification [88, 164], Drug Prescription [144, 145]
Business	Income Prediction [30, 97, 98, 170], Sentiment Analysis [5], Housing Prices Prediction [33, 117, 118], Banking [8], Investment [28, 43], Loan Approval [62], Profession Prediction [30, 133], Job Applications [29]
Educational	Nutrition [9, 10], LSAT [5], Kitchen Task [108], Student Performance Assessment [123], Analytical Reasoning (Text [27, 63], Spatial [22], Visual [114], Graph-based [40]), Spell Checking [40], Graduate Admissions [96]
Leisure	Speed Dating [94, 168], Block Completion [23], Bird Classification [18, 103, 131], Maze Completion [151], Board Game [114, 153], Penguin Identification [22], News Classification [112], Trip Planning [127], Logic Puzzles [144, 145], Exercise Recommendation [11], Movie Reviews [93]
Legal	Recidivism [32, 55, 156]
Science	Statelite Image Classification [18, 102], Leaf Classification [167], Forest Cover [156]
General	Hotel Review Classification [18, 61, 83, 132], Hand-written Digits Identification [27], General Image Recognition [146]

Table 7: A summary of intervention mechanics (datasets, data class, ML model, and deep learning models), decision-making types (binary, multi-class, and others), and target population type. There is a diverse use of datasets, while mostly studies use text/tabular or image/video datasets. Studies also use a mix of actual and simulated machine learning and deep learning models. Most interventions test binary decisions, while a handful use multi-class decisions. The majority of studies recruit crowd-workers and general participants.

Category	Studies
Dataset Type	Dating Profiles [94, 168], COMPAS [32, 55, 156], Income [30, 97, 98, 170], Food Ingredients [9, 10], Leaf [167], Forest [156], Reviews and Text Classification [5, 18, 27, 40, 61, 63, 83, 93, 112, 132, 133], Videos [108], Housing [33, 117, 118], Banking NLP [8], Investment [28, 43], Students [96, 123], Games [153], Loan [60, 62], Medical (Images, X-Rays, ECG, MRI, Profiles, etc.) [13, 17, 24, 71, 76, 88, 144, 145, 153, 164], Natural or General Images [18, 22, 27, 102], Birds [18, 103, 131], ImageNet [146], Job Applications [29]
Data Class	Text or Tabular [5, 8–11, 18, 27–30, 32, 33, 40, 43, 55, 60–63, 76, 83, 93, 94, 96–98, 112, 117, 118, 123, 127, 132, 133, 144, 145, 156, 164, 167, 168, 170], Images or Video [13, 15, 17, 18, 22, 24, 27, 71, 88, 102, 103, 108, 131, 146, 153], Patterns or Maze [23, 114, 151]
ML Models	SVM [83, 132, 167, 168], Random Forest [28, 168], Logistic/Linear/OLS Regression [33, 55, 62, 96–98, 117, 118, 123, 156], Gradient Boosting [29, 43, 60, 170]
Deep Learning	RoBERTa [5, 93], GoogleNet [108], MLP [8], VGG-19 [146], ResNet50 [131], ResNet18 [24], ResNext50 [15], ResNet [71], LogiFormer [63], CNN [27], FFNN [88], BERT [61]
Decision-Making	Binary [5, 9, 10, 13, 15, 17, 18, 22, 24, 27–30, 32, 40, 55, 60–62, 71, 76, 83, 88, 93, 94, 96–98, 102, 112, 114, 118, 123, 132, 133, 153, 156, 164, 167, 168], Multi-Class [5, 8, 11, 13, 18, 27, 30, 43, 63, 103, 108, 127, 131, 144–146, 151], Value-based [13, 33, 117], Pattern [23]
Participants	Potential Users [13, 15, 17, 88, 164], Crowd-workers [5, 8–11, 18, 27–29, 33, 40, 55, 60–63, 76, 83, 93, 94, 96–98, 102, 103, 112, 117, 123, 127, 131–133, 144–146, 151, 153, 156, 168, 170], General Participants [9, 22–24, 30, 43, 71, 88, 108, 114, 118, 167]

Table 8: An overall summary of included studies. Studies mostly use empirical evaluations with the crowd-worker population. While a few studies use single-stage decisions, a majority favor the multi-stage decisions. Studies mostly explore binary or multi-class decisions. Diverse explanation methods are used with a high focus on example-based and feature-based XAI. Notes: Most reported attributes are mutually non-exclusive. The dot (•) shows inclusion.

			Study Type			Target			Decisions			Tasks				Explanations				-
Sr	Study Ref	Year	Empirical	Qualitative	Mixed	Users	Crowd-workers	General	Single-stage	Multi-stage	Binary	Multi-class	Value	Other		Model-based	Feature-based	Example-based	General	Feedback
1	Yin et al. [168]	2019	•				•			•	•					•				•
2	Grgić-Hlača et al. [55]	2019	•				•			•	•					•				•
3	Lai and Tan [83]	2019	•				•		•		•						•	•		
4	Zhang et al. [170]	2020	•				•			•							•			
5	Buçinca et al. [9]	2020	•				•	•	•		•							•	•	
6	Yang et al. [167]	2020	•					•	•		•							•		•
7	Bansal et al. [5]	2021			•		•		•		•	•					•			•
8	Buçinca et al. [10]	2021	•				•			•	•						•			
9	Nourani et al. [108]	2021	•					•	•			•						•		
10	Lu and Yin [94]	2021	•				•			•	•						•			•
11	Wang and Yin [156]	2021	•				•			•	•						•	•		
12	Poursabzi-Sangdeh et al. [117]	2021	•				•			•			•						•	•
13	Cao and Huang [23]	2022	•					•		•				Pattern				•		
14	Chiang and Yin [33]	2022	•				•			•			•			•				•
15	Brachman et al. [8]	2022			•		•		•			•				•				
16	Dikmen and Burns [43]	2022	•					•	•			•						•		
17	Rastogi et al. [123]	2022	•				•		•		•								•	
18	Tejeda et al. [146]	2022			•		•		•	•		•						•		•
19	Schemmer et al. [132]	2023	•				•			•	•						•	•		
20	Schemmer et al. [131]	2023	•				•			•		•						•		
21	Cabitza et al. [13]	2023	•			•				•	•	•	•				•			
22	Vasconcelos et al. [151]	2023	•				•			•		•						•		
23	Wysocki et al. [164]	2023			•	•			•		•					•	•			
24	Vered et al. [153]	2023	•				•		•		•					•	•	•		
25	He et al. [63]	2023	•				•			•		•					•			
26	He et al. [62]	2023	•				•			•	•					•				
27	Chen et al. [30]	2023			•			•	•		•	•					•	•		
28	Prabhudesai et al. [118]	2023		•				•	•		•							•		•
29	Cabrera et al. [18]	2023	•				•		•		•	•				•				
30	Cau et al. [27]	2023	•				•			•	•	•					•	•		
31	Cau et al. [28]	2023	•				•		•		•						•	•		
32	Morrison et al. [102]	2023			•		•			•	•								•	
33	Cao et al. [22]	2023	•					•		•	•								•	
34	Chiang et al. [32]	2023	•							•	•								•	
35	Lee and Chew [88]	2023	•			•		•	•		•						•	•		
36	Ma et al. [97]	2023			•		•		•	•	•						•			
37	Peng et al. [114]	2024	•					•		•	•					•				
38	Cabitza et al. [17]	2024	•			•				•	•							•		
39	Cao et al. [24]	2024	•					•		•	•					•	•			
40	Ma et al. [98]	2024	•				•			•	•					•	•			
41	Pareek et al. [112]	2024			•		•			•	•								•	
42	Salimzadeh et al. [127]	2024	•				•			•		•							•	
43	Schoeffer et al. [133]	2024	•				•		•		•						•			
44	He et al. [61]	2024	•				•			•	•						•			
45	Morrison et al. [103]	2024			•		•			•		•						•	•	
46	Swaroop et al. [144]	2024	•				•		•	•		•							•	
47	Kim et al. [76]	2024			•		•			•	•					•				
48	Lu et al. [93]	2024	•				•			•	•								•	
49	Cabitza et al. [15]	2025	•			•				•	•						•	•		
50	de Jong et al. [40]	2025	•				•			•	•							•		
51	Jeon et al. [71]	2025	•					•		•	•						•	•		
52	He et al. [60]	2025	•				•			•	•						•			
53	Swaroop et al. [145]	2025	•				•		•	•		•							•	
54	Buçinca et al. [11]	2025	•				•			•		•							•	
55	Cau and Spano [29]	2025	•				•			•	•					•	•			
56	Ma et al. [96]	2025			•		•			•	•						•			
	Total	56	44	1	11	5	40	12	20	40	40	17	3	0		13	25	22	12	9