A Survey of AI Reliance

SVEN ECKHARDT, University of Zurich, Switzerland

NIKLAS KÜHL, University of Bayreuth & Fraunhofer FIT, Germany

MATEUSZ DOLATA, University of Zurich, Switzerland

GERHARD SCHWABE, University of Zurich, Switzerland

Artificial intelligence (AI) systems have become an indispensable component of modern technology. However, research on human behavioral responses is lagging behind, i.e., the research into human reliance on AI advice (AI reliance). Current shortcomings in the literature include the unclear influences on AI reliance, lack of external validity, conflicting approaches to measuring reliance, and disregard for a change in reliance over time. Promising avenues for future research include reliance on generative AI output and reliance in multi-user situations. In conclusion, we present a morphological box that serves as a guide for research on AI reliance.

CCS Concepts: • **Human-centered computing**;

Additional Key Words and Phrases: Artificial intelligence, Reliance, Literature survey

1 INTRODUCTION

The term *Artificial intelligence (AI)* was first proposed at the Dartmouth Conference in 1956 [87]. Since then, considerable progress has been made toward the development of modern AI systems, such as transformer models [126] capable of human-like speech [51] or image generation [70]. This rapid performance growth has led to an increase in the use of AI-based solutions in productive environments and the democratization of AI access. One example of the pervasive use of AI can be observed in the publication of GPT models to everyday users through the interface of ChatGPT [1]. This illustrates the trend of the general public accessing and using cutting-edge systems.

The ChatGPT example also highlights an issue with the widespread publication of AI systems. While there has been research on the underlying large language model (LLM) *prior* to its release, there has been little research on the interaction between its users and the tool. To gain a better understanding of the interaction between users and ChatGPT, it was necessary to conduct research only after the tool had been made available to the general public. This meant that the tool's potential positive and negative effects, becoming apparent in humans' interaction with them, were—and still are—not fully understood. In light of this, it is becoming increasingly clear that more regulations are needed to govern

Authors' addresses: Sven Eckhardt, eckhardt@ifi.uzh.ch, University of Zurich, Zurich, Switzerland; Niklas Kühl, kuehl@uni-bayreuth.de, University of Bayreuth & Fraunhofer FIT, Bayreuth, Germany; Mateusz Dolata, dolata@ifi.uzh.ch, University of Zurich, Zurich, Switzerland; Gerhard Schwabe, schwabe@ifi.uzh.ch, University of Zurich, Zurich, Switzerland.

the use and development of AI systems. A noteworthy example is the EU AI Act [45], which aspires to be *"the world's first comprehensive AI law"* with the objective of regulating the development and use of AI systems.

The full implications of human *behavior* with regard to these novel AI systems remain poorly understood. It is possible that individuals may use these systems to improve their performance or as lazy shortcuts, as evidenced by reduced cognitive effort when confronted with AI advice [47]. This shows that not only the performance of AI systems is important, but also the reliance of users serving as decision-makers on AI advice—a phenomenon we call *AI reliance*. Overall, it is important to consider not only the technical capabilities of AI systems, but also a sociotechnical view of the AI system and the human decision-makers for a productive use of AI systems.

The pervasive use of AI in research and practice has led to a growing interest in the topic of AI reliance in order to understand how users behave when confronted with AI systems and their recommendations. If a user follows the AI advice, we can infer that the user *relies* on that advice. Conversely, if a user follows incorrect advice, we can infer that the user is *overrelying* on the AI advice. If a user does not follow correct advice, we can infer that the user is *underrelying* on the AI advice. Only by taking into account AI reliance can we investigate whether people are using the systems appropriately or blindly following them, therefore overrelying on them—or potentially ignoring their recommendations entirely.

An illustrative example of the role of AI reliance is the case of the COMPAS [18] system, an AI system designed to classify the recidivism risk of criminal defendants. In this case, instances of overreliance and underreliance have a significant impact on individual defendants, who may be sentenced to longer prison terms, and on society in general, which is confronted with individuals who have a high likelihood of reoffending. This example highlights the core challenge of AI reliance in practice: achieving *appropriate reliance*. The human-AI team can then achieve a complementary performance that surpasses that of either the human or the AI alone [7]. This increases the overall performance of the human-AI collaboration compared to that of each individual component.

Despite the overall importance of AI reliance, there is no clear guidance on how to conduct AI reliance research. For instance, researchers might face multiple inconsistent definitions and measures, of which some remain niche. There are several theoretical and conceptual considerations around AI reliance[e.g. 69, 104, 119]. These considerations mostly aim at differentiating trust and reliance and base their conclusions on the philosophical literature around reliance in a more general sense. A clear focus on *AI* reliance is missing. Whereas some recent approaches aim to formulate a *formal definition of reliance* [52], they focus solely on the definition of AI reliance and look at how reliance has been established in past research. Consequently, while they offer a theoretically sound conceptualization, this might not reflect how AI reliance has been understood in research. Accordingly, researchers who want to engage with this topic might find themselves torn between theoretical perspectives and their peers' research practice. Researchers who are studying AI reliance are still navigating uncharted waters in their efforts to identify the most appropriate methodology and fruitful avenues for conducting AI reliance research. Overall, the current state of AI research can be described as rather chaotic and unstructured, with a variety of definitions, measures, conceptualizations, and general understandings. In the long run, this may result in the replication of unsound practices or the application of research methods without an understanding of their implications. A systematic understanding of researching AI reliance is therefore important, and is presented in this review.

Review Synopsis

This survey is structured as follows. In Section 2, we provide the background for this survey, particularly on AI reliance. We also describe the notion of a sociotechnical system, which we use as a fundamental perspective for analyzing

literature on AI reliance. In Section 3, we describe the methodology used to gain insight into the literature. We employ a structured approach to query existing literature and derive a well-founded set of relevant articles, including keyword search and forward and backward searches. This enables a well-founded set of articles to be analyzed. Afterwards, in Section 4, we provide a categorization framework that is used to analyze the relevant literature and can assist for future researchers to conduct well-founded studies on AI reliance. In Section 5, we analyze the current literature on AI reliance using this categorization framework, which enables an in-depth and systematic investigation of the literature. Finally, based on this review, in Section 6 we discuss the literature and present future research avenues, and conclude this survey with a brief summary in Section 7.

This review offers a comprehensive and systematic understanding of the current research on AI reliance and makes several significant contributions to the field. First, it analyzes current approaches to research reliance and provides an overview of the scattered field of AI reliance. It introduces a sociotechnical perspective on AI reliance, which is essential for a thorough understanding of this topic. Second, it discusses the limitations of the current literature on AI reliance. It identifies areas that require further investigation in future research on AI reliance. Third, we identify new research avenues that require attention in future studies on AI reliance by presenting emerging issues. Finally, we present a categorization framework that researchers can use as a template for conducting well-founded AI reliance studies. This framework ends up in a morphological box that can serve as a template to present AI reliance research.

This survey provides substantial support for two distinct groups of AI reliance researchers. The first group is interested in extending the field of AI reliance. In this case, the survey presents current limitations and suggests future avenues for conducting research on this topic. The other group is interested in fostering appropriate reliance on their AI systems and presenting their results. In this case, the survey provides an overview of current approaches to AI reliance and, more importantly, a framework to use when conducting AI reliance research.

2 BACKGROUND ON AI RELIANCE

2.1 Related Literature Surveys

The field of AI reliance is still in its infancy, with no unified, established, and broadly accepted conceptualizations or definitions. It is however evident that research on AI reliance is closely related to several other disciplines in the domain of human-centered AI systems. In this context, we will briefly introduce literature surveys on related concepts and the relationship between these and AI reliance.

Once modern AI systems based on machine learning achieved sufficient performance to be used in productive settings, practitioners and researchers began to focus on the design aspects of these systems to achieve human-centered AI systems. Most research in this field has focused on the explainability of AI systems. One of the earliest reviews in this area was conducted by Adadi and Berrada [2] in 2018, who were among the first to systematically structure the existing literature on explainable AI. More recently, Dwivedi et al. [41] published a comprehensive overview of the core ideas, techniques, and solutions associated with explainable AI. Another relevant topic is the fairness of AI systems. In general, it is desirable that AI systems output fair decisions that do not discriminate. There is however no single definition of fairness but rather several concepts of fairness, which for example Mehrabi et al. [89] or Pessach and Shmueli [100] surveyed. Up to now, there is no systematic consideration which correlates with AI reliance.

Explanations, fairness, and related concepts are often employed to achieve trustworthy AI systems, as surveyed in Kaur et al. [65]. They present one of the first sociotechnical considerations of the interplay between user and AI system, without explicitly stating the perspective. While the aforementioned studies focus on trust, this review will concentrate

on reliance behavior. With regard to human-AI decision-making, the survey by Lai et al. [75] represents one of the first attempts to examine the existing literature in a structured manner. While this survey is relevant to the present study, it however lacks a clear focus on reliance and only considers literature up until 2021.

Research on fair, responsible, and explainable AI has shown clearly that we need to frame complex phenomena related to AI as contextualized in the usage practice and organizational context. Many have proposed to follow the sociotechnical systems lens as an adequate perspective to frame the phenomena and understand research surrounding them [39, 73, 125]. They were shown to generate new insights and, particularly, define new avenues for future research. Whereas existing conceptual work and overviews focused on technical or theoretical definitions [e.g. 52], we follow the sociotechnical perspective to offer a comprehensive, overarching view of AI reliance, comprising studies from various subfields of computer science.

2.2 Background on AI Reliance

The notion of reliance is not unique to AI or computer systems—it is extensively discussed in the field of philosophy and psychology. Reliance is often aligned with trust, as the two concepts appear similar at first glance [6, 55]. However, many philosophers argue that trust cannot be expressed toward inanimate objects [e.g. 55, 58]. Nevertheless, a substantial body of AI research has been conducted on trust in AI, often referred to as trustworthy AI [65]. It can be argued that computer systems (and AI) are not inanimate objects, as evidenced by the "computers are social actors" principle [95]. When computer systems research discusses trust, it is however often concerned with reliance [36]. In this article, we follow the line of discourse that suggests differentiating between trust (an emotional or attitudinal stance toward something) and reliance (observable behavior) [e.g. 78]. This review focuses on studies that investigate behavior rather than attitude. The remainder of this section describes various discourses of human reliance on computer systems. We describe how reliance on automation systems has influenced research about reliance on intelligent systems. We also demonstrate how findings from the discipline of advice-taking influence research about reliance on AI to create so-called judge-advisor systems.

Automation Systems. The discussion of the reliance on (intelligent) computer systems commenced with automation systems [116]. Automation is defined as *"the execution by a machine agent (usually a computer) of a function that was previously carried out by a human"* [99, p. 231]. This notion is evolving over time with the capabilities of computer systems. In this survey, we differentiate automation systems from AI by their focus on automating tasks and lack of predictive power. We also acknowledge that this distinction is becoming increasingly blurred. A simplified example of an automation system where the topic of reliance becomes important is a conveyor belt with a warning light that turns on once there is overheating. In that example, reliance can be defined as the human intervening as soon as the light turns on.

Previous research has frequently concentrated on the concept of *trust* in human-machine interaction in automation systems [e.g. 92, 93]. A frequently researched case is that of pilot cockpit automation systems [e.g. 22, 129]. When reading these earlier studies, it becomes evident that the concept of trust is more complex than previously thought. For example, Wickens [129, p. 366] defines trust as *"the extent to which the pilot believes that, and **behaves** as if the automation will carry out its assigned task in a reliable fashion."* This definition encompasses not only the pilot's attitude toward the system but also their behavior.

This distinction between attitude and *use* has been further considered by Parasuraman and Riley [99], who define the *use*, *misuse*, *disuse*, and *abuse* of automation systems. They define *use* as *"the voluntary activation or disengagement of*

automation by human operators," misuse as the *"overreliance on automation,"* disuse as the *"neglect or underutilization of automation,"* and abuse as the *"automation of functions by designers and implementation by managers without due regard for the consequences for human performance"* that might also resolve in the misuse and disuse of the system [99, p. 230].

While this early definition of what is essentially reliance remains descriptive of the possible behavior of humans, we can already derive the desired behavior for human operators, which is, appropriate reliance [42, 78]. To achieve this, the human operator uses the automation system if it is correct, but does not use it if the system is incorrect. This is neither a misuse (overreliance) nor a disuse (underreliance) of the automation system.

Intelligent Systems. In addition to the automating processes humans previously performed, computer systems can also serve as decision aids. We use intelligent systems for many similar notions, such as (intelligent) decision support systems. AI is often seen as the foundation for developing intelligent systems [e.g. 108], most often enabled by machine learning methods [48]. AI use gives these systems the ability to predict outcomes and distinguishes them from automation systems. In the case of the COMPAS algorithm [18], the objective was to predict whether an offender would reoffend after being released. This (AI) system provided advice to a human judge, who made the decision after being presented with the AI advice.

Some research about reliance on AI systems has followed the early notions of reliance on automation systems and defined reliance as *the following of AI advice* [78]. To continue the running example of COMPAS, a judge is said to rely on the system if the release of offenders with low recidivism risk for jail is early but retains high-risk offenders. This reliance, however, does not indicate whether this was the correct decision or not. It is often the case that AI systems are probabilistic in nature and can therefore provide erroneous advice. Consequently, the concepts of underreliance and overreliance become important. Overreliance (*misuse* of intelligent systems) refers to the acceptance of incorrect advice. Conversely, underreliance (*disuse* of intelligent systems) refers to the rejection of correct advice. In a manner analogous to the definitions of automation systems, appropriate reliance can be defined as the acceptance of correct AI advice and the rejection of incorrect AI advice.

This understanding of reliance can be measured and quantified by counting the number of times a user follows the AI advice. In the literature on AI reliance, this is often referred to as *the agreement fraction* or *agreement percentage.* [e.g. 76, 124]. With this understanding, it is possible to measure underreliance (the number of instances where the user does not follow the AI's correct advice), overreliance (the number of instances where the user follows the AI advice despite it being incorrect), and appropriate reliance (the number of instances where the user follows the AI's correct advice and does not follow incorrect advice). Many more metrics are used in literature, such as the weight of advice, which is introduced in the following section.

Judge-Advisor Systems. Another angle of reliance on AI systems is to lend insights from advice-taking literature. This might be necessary as presenting advice may not be exactly the same as automating processes and turning on a warning light in the example of a conveyor belt. In advice-taking literature, the concept of the judge-advisor system (JAS) is widespread [118]. In a JAS, the judge gets a task and receives advice from an advisor. Based on this task and advice, the judge makes the final decision. This JAS also introduces a new angle on advice-taking, where the judge gives an initial estimate without being exposed to the advice. Then an advisor is employed to advise the judge on their decision, whereafter the judge can give a final decision. In the case of the COMPAS algorithm [18], the judge would first have to make their own assessment. Only then the advice of the COMPAS system would be presented to the judge, who then makes the final decision.

This JAS can be transferred to AI, where the AI serves as an advisor to the human (judge) by providing advice. This can provide a more granular view of AI reliance. Some research suggests that reliance can only be observed if the initial decision by the human differs from the AI advice [114]. This is because if they both coincide, it is not possible to determine whether the human relied on the AI advice or if this agreement occurred coincidentally. Overreliance is defined as a situation in which a human changes from a correct decision to an incorrect one after being exposed to incorrect advice. Underreliance is defined as a situation in which a human does not change from an incorrect decision to a correct one after being exposed to correct advice. Finally, appropriate reliance is defined as a situation in which a human does not change to an incorrect decision after being exposed to incorrect advice, or change to the correct decision after being exposed to correct advice.

This conceptualization of an initial human decision also permits a more comprehensive examination of reliance. Rather than merely counting instances of agreement with AI advice, some researchers are interested in the number of instances where the user alters their decision after being confronted with the AI advice. This phenomenon is often referred to as the *switch fraction* or the *switch percentage* [e.g. 115, 131]. Moreover, reliance can be quantified in the absence of discrete decision-making or even regression cases. This is derived from the literature on advice-taking, whereby the *weight of advice (WOA)* is calculated by: *(final estimate - initial estimate) / (advice - initial estimate)* [118].

Summary. The preceding discourses on human reliance on computer systems underscore the necessity of considering the impact of AI reliance from multifold perspectives, rather than from a single, narrow one. The perspective of a sociotechnical system can provide such a multifold perspective. In the following section, we introduce this perspective and show its relation to the topic of AI reliance.

2.3 Sociotechnical Systems

Sociotechnical systems consist of a *technical component* (e.g. the AI system), a *social component* (e.g. the human decision-maker), and an *interaction* between these two in a broader context or *environment* (e.g. an organization). When considering AI reliance, it is important to recognize that the final result of the interaction between the AI system and the user occurs in a specific environment and is contingent upon both the human decision and the AI advice. A sociotechnical lens is therefore essential. Figure 1 presents a simplified representation of the four components of a sociotechnical system and the relation between them. The individual components are introduced in the following section and used as a basis for the analysis of the literature on AI reliance.
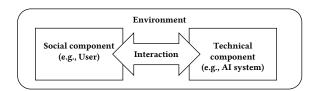

Fig. 1. Simplified representation of a sociotechnical system.

The technical component is defined as comprising technical and material artifacts and the techniques or practices employed to use the artifact [77, 123]. In the case of human-AI interactions, this would primarily be the design and advice type of the AI system. The social component encompasses individuals or collectives, as well as the relationships between them, which may be expressed as roles, hierarchies, structures, economic systems, cultures, power relations,

or communication networks [82, 110]. In the context of human-AI interaction, this would be the human user/decision-maker. The sociotechnical system is however more than the sum of the technical and social components. These components enter complex mutual interactions, therefore an understanding of the system's functioning is only possible if one considers the interdependencies between the components and their subcomponents [82]. An analysis that focuses on only one of the components offers an incomplete perspective of the system, which might limit insights for identifying, for example, the overall system outcomes. The interaction between the social and technical components is recursive and reciprocal, forming a process of joint optimization [110]. A sociotechnical system also operates in a broader societal, political, regulatory, organizational, and historical context [29]. It is essential to consider the interaction with the environment, which can be framed in terms of an input-transformation-output model. Consequently, a sociotechnical system receives inputs from the environment, such as economic pressure or higher demand, which it transforms to deliver the output that fits the environment.

2.4 Summary

The current research on AI reliance aims to understand and design for appropriate reliance on AI advice [e.g. 12]. The goal of this research is, typically, to achieve complementary team performance, where the team of humans and AI exceeds the performance of the human or AI components if considered separately [8]. There is no consistent approach across the related work. While there have been reviews analyzing human-AI decision-making [e.g. 75], there has been no focus on the reliance of decision-makers on AI advice. These studies frequently focused on a single aspect of the sociotechnical system, namely the technical component or the AI system itself. The shortcoming of this approach is exemplified by the case of explanations [e.g. 19, 114] that often show inconclusive results [113]. We argue that one reason is that a comprehensive perspective is missing. The sociotechnical system perspective enables us to introduce a comprehensive perspective and analyze existing literature on AI reliance structurally. This allows us to establish uniform approaches and provide guidance, which is essential for research to have a real-life impact and achieve the desired appropriate reliance on AI advice.

3 SURVEY METHODOLOGY

Given the lack of unified approaches in the literature, a structured literature review is employed to provide an overview of the current approaches researchers who investigate AI reliance employ. This section describes the methodology used for structurally querying the literature. Based on a classification framework introduced in Section 4, the results of the structural literature review are presented in Section 5. These results are then used in Section 6 to discuss the shortcomings and limitations of current approaches and to present guidance for researchers conducting sound AI reliance research. Additionally, future avenues for further developing the field of AI reliance are presented.

3.1 Queries

To identify relevant articles for research into AI reliance, we employed a structured query of scientific databases. As there are numerous underlying technologies for AI systems, we also queried the common terms that are treated as synonyms, such as machine learning and deep learning[1]. In order to include articles that explicitly investigate overreliance and underreliance, we added these terms to the search string, as indicated in Table 1.

---

[1]While we are aware of the technical distinction between artificial intelligence and machine learning as discussed, for example by [74], we acknowledge that the synonyms have established themselves in the computer science and human-computer interaction communities.

Queries	**(**Artificial Intelligence **OR** AI **OR** Machine Learning **OR** ML **OR** Deep Learning **OR** DL**)** **AND (**Reliance **OR** Overreliance **OR** Underreliance**)**
Databases	ACM Digital Library (https://dl.acm.org/), AIS eLibrary (AISeL) (https://aisel.aisnet.org/), IEEE Xplore (https://ieeexplore.ieee.org/), Scopus (https://www.scopus.com/)
Cutoff date	December 31, 2023

Table 1. The metadata for our search, including the keywords, the databases queried, and the cutoff date for the search.

The search was conducted on the articles' titles, abstracts, and author keywords. A less restrictive search for indexed keywords would result in the inclusion of articles that are not directly related to AI but are still indexed under the AI keyword, such as those published in outlets with AI in the name. This approach was not adopted, as it would result in the inclusion of irrelevant articles. In case of different spellings, the keyword reliance will also present a hit in the database for hyphenated compound words, such as "over-reliance," negating the necessity to include these words explicitly in the search string. We also considered including the terms "rely" and "relies," as some articles might not explicitly state the concept *reliance* in their title, abstract, or keywords. This would however lead to a much larger number of articles, the vast majority of which are not related to AI reliance. We therefore opted to only include reliance, to focus on articles that explicitly investigate that concept.

Four databases were queried, with the cutoff date for the selected article set to 31 December 2023. The databases are summarized in Table 1 and include the *ACM Digital Library*, *AISeL*, *IEEE Xplore*, and *Scopus*. The rationale is that the first three collectively represent the most important outlets in the categories of computer science, human-computer interaction, and information systems, while SCOPUS provides a broader view of existing literature. Given the vast scale of SCOPUS, we focused only on the top 25 % of outlets according to their citation score in the categories of *computer science* (including AI and IS), *business*, *management and accounting*, *decision sciences*, *multidisciplinary*, *psychology*, *general social sciences*, and *human factors and ergonomics*. Articles that were not initially identified due to this focus will be identified in the subsequent forward and backward searches (Section 3.3)

3.2 Inclusion and Exclusion Criteria

The query yielded 1,337 articles, but not all of them concerned the topic of AI reliance. To determine the relevance of each article to this survey, we defined certain inclusion and exclusion criteria, summarized in Table 2 and introduced in the following section.

First, as an obvious inclusion criterion, we included articles that used the abbreviations AI, ML, and DL in the search string. Some articles employed these abbreviations in a manner that differed from the anticipated usage. For instance, AI was used to refer to "American Indian," DL to "discrete logarithm" or "dictionary learning." Articles that deviated from the anticipated use of the abbreviations and did not refer to artificial intelligence, machine learning, or deep learning in any other form were excluded. Most importantly, only articles where reliance is considered in the context of humans relying on AI systems were included. Many of the articles have a clear technical perspective, with the term "reliance" mostly used to refer to the reliance of AI on datasets, labeled data, or similar resources. Given the extensive scope of the articles and disciplines under consideration, the term "reliance" was also employed to describe reliance on objects or entities, such as "reliance on agriculture" or "the world's reliance on Chinese exports." These latter examples were

#	Inclusion criteria	Exclusion criteria
1	Articles that use the abbreviations AI, ML, or DL as intended in this study	Articles that use the abbreviations AI, ML, or DL differently
2	Articles concerned with reliance on AI (or ML/DL) systems	Articles that use reliance in other contexts, such as the reliance of AI systems on data
3	Articles that examine the user behavior that occurs when individuals interact with AI systems	Articles that focus on the philosophical aspects of reliance on AI in general, maintaining an abstract level of discourse
4	Articles in English	Articles in languages other than English
5	Articles published in or after 2010	Articles published before 2010
6	Archival peer-reviewed articles	Other articles, such as editorials or preprints

Table 2. Inclusion and exclusion criteria for the query (#1-3) and eligibility criteria (#4-6)

excluded from this study. Finally, it became evident that the aforementioned criteria also encompassed articles that perpetuate the philosophical discourse on reliance by introducing AI systems, such as the differentiation between trust and reliance. These articles however fail to consider the actual and individual AI reliance, meaning the user behavior when confronted with AI advice. Consequently, this study focuses on articles that research AI reliance by investigating human behavior, and articles of a purely conceptual nature are included in the study's related work.

In addition to the aforementioned inclusion and exclusion criteria, we employed a set of general eligibility criteria. These criteria relate to the type of article that we include and exclude. Specifically, we only included full articles and exclude editorials, short articles, extended abstracts, presentations, preprints, and similar. Furthermore, we only considered articles in English that were published after 2010. The year was selected as the cutoff point for the query because the field of AI research underwent significant changes during the 2010s, with increasing computing capabilities and data availability. These changes were reflected in the emergence of modern deep learning and neural networks, which only began to gain prominence around that time.

After applying the inclusion and exclusion criteria, we had 45 relevant articles. We conducted a forward and backward search to ensure that no relevant articles were missed due to a narrow keyword search or a narrow databased selection.

3.3 Forward and Backward Search

A forward and backward search was conducted after selecting relevant articles based on querying the literature databases. All identified relevant articles were used for the forward search. This search was performed on SCOPUS without any additional filters[2].The forward search yielded an additional 441 articles, and the backward search an additional 1,160 articles. In total, we found another 26 relevant articles. Two main reasons account for the absence of these articles in the initial search. First, some articles were published in an outlet not included in the top 25% of SCOPUS. Second, some of these articles did not mention reliance in the title, abstract, or keywords, but rather closely related concepts, such as algorithm aversion or automation bias, and only discussed reliance in the main body text.
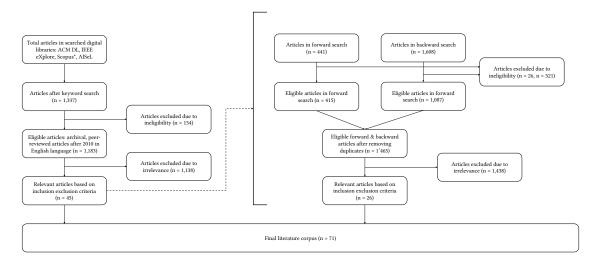
Fig. 2. PRISMA flow chart [90] of queried literature [3, 5, 8, 10, 11, 13–15, 17, 19–21, 23–28, 30–33, 35, 37, 38, 40, 43, 46, 49, 50, 53, 56, 57, 59, 61, 62, 66, 71, 76, 79–81, 84–86, 88, 91, 94, 96–98, 101–103, 105–107, 111, 112, 114, 115, 120–122, 124, 127, 128, 130–133]

3.4 Query Results

In the following section we summarize the query results. A total of 71 relevant articles were identified through the application of the described methodologies. The query results are summarized in Figure 2 using a PRISMA flow chart [90].

The initial database search yielded 1,337 articles, of which 154 were excluded due to them not being in English, having been published before 2010, or not having been peer-reviewed, leaving 1,183 articles. After screening the articles using the inclusion and exclusion criteria, we identified 45 relevant articles that investigated the empirical use of AI.

We then conducted a forward and backward search, which yielded 441 unique articles in the forward search and 1,608 unique articles in the backward search. After excluding ineligible articles, we were left with 415 articles for the forward search and 1,087 articles for the backward search. As some articles were included in both the forward and backward search sets, merging the two yielded 1,465 unique articles. After screening these articles concerning the inclusion and exclusion criteria, we were left with 26 additional articles that concerned the empirical investigation of AI reliance.

This resulted in 71 unique articles dealing with the empirical study of AI reliance. The results are summarized in Figure 2. The examination of basic bibliometrics, such as publication year or outlet, provides insight into the communities concerned with AI reliance. This is described in the following section.

3.5 Bibliometrics

Figure 3 shows a histogram of the publication dates. Despite our search including articles from 2010 onwards, the earliest articles identified are from 2018[3]. It is evident that the number of published articles has increased significantly
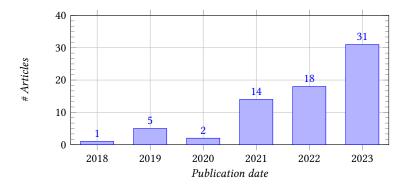

Fig. 3. Histogram of the identified relevant articles' publication dates.

since 2021. This suggests that research on AI reliance commenced at around this time. The EU AI Act was also first discussed in 2021 [44], demonstrating the timeliness and relevance of the topic of AI reliance.

Source	Articles
Proceedings of the ACM on Human-Computer Interaction (PACMHCI)	16
Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)	8
Proceedings of the International Conference on Intelligent User Interfaces (IUI)	6
ACM International Conference Proceeding Series (ICPS)	4
Proceedings of the International Conference on Information Systems (ICIS)	3

Table 3. The top five most relevant sources for the selected articles.

Table 3 presents the five most relevant sources from which the articles originate. As might be expected, the HCI literature is the subject of considerable attention, with outlets such as the PACMHCI (and its individual conferences) and CHI being listed in the ACM Digital Library. Furthermore, one of the top five sources is the AISeL (ICIS). It is noteworthy that the IEEE Xplore database does not yield a substantial number of relevant articles, either in the list of the top five most relevant sources for the relevant articles or in the set of final relevant articles. This indicates that the topic of AI reliance is currently most dominantly centered around HCI research, with only a slight touch upon core computer science research.

4   CLASSIFICATION FRAMEWORK FOR THE LITERATURE

To gain an understanding of the existing literature, the identified articles are classified according to a range of concepts. These concepts are based on the components of a sociotechnical system and the subconcepts derived in an author workshop based on a set of preselected articles. This filtering and classification process is then employed to gain insights into how the existing literature conceptualizes AI reliance. The steps are explained below.

The sociotechnical system provides a profound perspective to consider AI reliance. As described in Section 2, a sociotechnical system consists of four components: (a) the *environment*, (b) the *interaction*, (c) the *social component*, and (d) the *technical component* in which the system is situated. These components serve as the basis for classifying the relevant articles based on their primary focus. To derive the subconcepts in the four groups, three authors held an author workshop. Prior to the workshop, each author read a subset of the articles, focusing on common themes,

Concept	Subconcept	Description
(a) Environment	(1) *Task*	Humans are confronted with a variety of decision problems. One can distinguish between decision problems with objectively correct advice, such as detecting diseases in X-ray images, and decision problems with no objectively correct advice and only subjective results, such as music preferences.
	(2) *Setting*	The manner in which data is collected for articles may vary, and this has implications for the results. Literature can be classified according to the manner in which data is collected, such as from crowdworkers, like MTurk, or from domain experts, such as medical experts.
	(3) *Use cases*	Humans can receive advice in a multitude of contexts. Consequently, the present literature can be classified according to the specific use cases where an AI system is used to present advice to a human, such as in the financial or medical domain.
(b) Interaction	(4) *Decision-making approach*	There are multiple ways in which AI advice can be presented to a human decision-maker for incorporation into their decision-making approach. A distinction can be made between a single-stage approach, where AI advice is directly communicated to the user, and a two-stage approach, where the user must make an initial decision without the advice.
	(5) *Reliance measure*	To quantify the interaction between the social and technical components, it is necessary to measure the extent to which humans rely on AI advice. This reliance can be measured using a variety of metrics that have been proposed in the literature.
(c) Social	(6) *User training*	AI systems are often complex in nature, and users may not immediately understand the implications and derivation of advice. Initial training could help users understand how the AI system works, and influence reliance. However, not all articles employ training, therefore this can be used as a way to categorize the literature and distinguish between articles with and without training.
	(7) *Performance feedback*	It is common for humans to interact with AI systems on multiple occasions. This is also the case in the majority of current research, where reliance is considered over multiple task instances. Consequently, one can distinguish between articles that do not provide feedback during these instances and those that provide some feedback after a decision task. This could help users calibrate their reliance on AI advice.
(d) Technical	(8) *AI system implementation*	The implementation of AI systems is typically accomplished through various algorithms and architectures. In the context of user studies, users may not necessarily interact with a live AI system. The existing literature on this topic can be broadly classified into three categories: WOZ approaches, articles that manually sample the decision instances communicated to the user, and articles where users are confronted with the live AI systems.
	(9) *Transparency mechanism*	A common design choice for AI systems is the incorporation of transparency mechanisms. Consequently, articles can be classified into three categories: those that present explanations to the user on how the AI system arrived at the advice, those that present a form of performance/uncertainty of the AI system, and those that do not present any transparency mechanism to the user.

Table 4. Concepts used to classify the literature on AI reliance, based on the four components of a sociotechnical system: (a) environment, (b) mutual interaction, (c) social component, and (d) technical component.

concepts, and interesting findings. The concepts were derived and subsequently refined during the workshop. Nine subconcepts, distributed across the four components of a sociotechnical system, were identified as relevant to the study and agreed upon by all authors. These concepts are used to classify and describe the articles and are summarized in Table 4.

For the component of the environment, we identified three subconcepts: (1) *task* states whether the decision problem at hand has objectively correct advice, meaning a clear ground truth, or whether the AI should present advice for a decision task with a subjective outcome. Examples of the latter category include ethical dilemmas such as the trolley problem, where a decision-maker must determine whether to sacrifice one individual in order to save a larger number of people. Another example is personal preferences, such as selecting a song based on mood or music taste. To add to this, the (2) *setting* in which the data is collected is also important, as it may influence the reliance of human decision-makers. For example, whether the data is collected in a real-world setting or in a laboratory experiment might have an impact on user behavior. Finally, the articles can be differentiated according to their (3) *use cases*. This should provide a general overview of the coverage of AI reliance research in real-world problems. This may also allow for the identification of domains where AI reliance is considered more often than others.

In addition to the environment, the sociotechnical system also consists of the mutual interaction between the social and technical components. As interaction, we consider the reliance of the social component on the technical component, meaning the reliance of the user on the AI system. This is influenced by the (4) *decision-making approach* used in the different articles. In general, there are two possible approaches. On the one hand, there is the one-step decision-making approach, where the AI advice is directly communicated to the user at the same time as the decision task. On the other hand, there is the two-step approach, where the human has to make their own initial assessment before revising it after being confronted with the AI advice. These approaches may also influence the (5) *reliance measurement*. Only with unified measurement are results comparable and statements generalizable. However, a variety of measures have been used to assess reliance, and there seems to be no consistent approach. We therefore examine how the articles measure reliance. This should provide insight into how the articles address the interdependence between social and technical components.

The social component relates to the user or decision-maker who either relies or does not rely on the AI advice. The user can calibrate the reliance based on the experience they have with the tasks and the AI system. On the one hand, a (6) *user training* can be conducted before the decision-making tasks, for instance through trials and tutorials. This helps the user get familiar with the AI advice. On the other hand, training can also be continuous (and live) during the tasks. This involves providing (7) *performance feedback* about the decisions based on the AI advice. Both approaches are considered when analyzing the literature. Both have the potential to influence the reliance of users on the system and therefore the sociotechnical system's social component.

The technical component refers to the AI system itself, therefore it is important to cover relevant aspects of the system. Consequently, it is important to evaluate the (8) *AI system implementation*, as this influences the reliance. We abstract this performance to the three cases of "Wizard of Oz" (WOZ) studies, manually selected labels by the authors, and live AI system. In WOZ studies, no genuine AI system is constructed; rather, users interact with human-generated advice, which is labelled as AI advice. One of the most common design decisions for AI systems, especially when user interaction is a primary concern, is the incorporation of (9) *transparency mechanisms*, such as explanations. By examining these subconcepts, we can gain a comprehensive understanding of the technical component.

The concept is based on the components of a sociotechnical system, which provides a perspective for examining AI reliance. The subconcepts were derived by gaining insight into the articles and extracting the most important
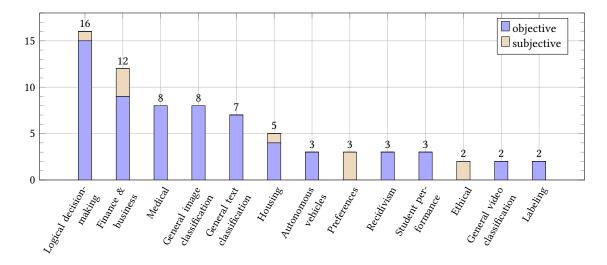
Fig. 4. Overview of all use case categories and categorization of whether the goal of the task has no objective correct advice or whether advice can only be subjective. Some articles present tasks from multiple categories and are therefore listed multiple times.

and interesting aspects. These concepts and categorizations provide a framework for literature on AI reliance. In the following section, we present the results of categorizing the literature on these concepts and subconcepts, providing a structured overview of the current literature.

5 RESULTS

Based on the categorization framework introduced in the previous section (Section 4), we analyze the current literature on AI reliance in this section. In Section 5.1, we present the environment where AI reliance occurs in the identified articles. In Section 5.2, we present the general conceptualizations of the mutual interaction of the technical and social components, namely reliance, by reviewing the measures and decision-making approaches. In Section 5.3, we address the social component of the sociotechnical system and review the way users come to know the systems. Finally, to include the technical component, we review the properties of the AI system in Section 5.4.

5.1 Exploring the Environment of AI Reliance

The first component of the categorization framework is the environment in which AI reliance is observed. This is based on the subconcepts of the task, the setting, and the use cases in which AI reliance is studied. It is notable that, despite the prevalence of real-world use cases in daily life and work, most articles employ crowdworker marketplaces, such as Amazon Mechanical Turk (MTurk) or Prolific, instead of professionals or experts. Indeed, 47 articles use online crowdworker marketplaces to recruit participants. The remaining articles employ a variety of recruitment strategies, including domain experts (n = 9), students (n = 5), and convenience sampling (n = 10), without further restrictions on the participants. This indicates that most articles use crowdworker marketplaces to recruit participants, and nearly all articles examined reliance in a controlled and isolated setting.

A multitude of use cases are employed and Figure 4 presents an overview. For a more in-depth analysis, we can divide the use cases into two groups: those where the AI aims to present objectively correct advice, meaning use cases

with a ground truth present, and those where the outcome is of a subjective nature, linked to user preference. Both require fundamentally different AI systems, with different goals. One system can be trained with an objective ground truth, while the other systems aim to estimate the individual user preferences. It is important to distinguish these cases and become aware of the specific use case that the AI system aims to address.

A prime example of the group of subjective use cases is recommender systems, such as music recommendations [84, 103] or recommendations on the attractiveness of other people [106]. In these cases, the AI system cannot predict an objective ground truth (because there is none) but must rather aim at the user's preferences. Besides recommendations, AI can also assist with creative tasks [59]. Ethical decision-making is also often seen as having subjective outcomes. Examples from the AI reliance literature are ethical decision-making tasks, such as ethical dilemmas. Examples in the literature are military defense and rescue actions [122] or a decision about a recipient for a donor kidney, where multiple viable options exist [94]. Subjective outcomes can also be found in use cases where one might not expect them at the outset. For example, managers may have preferences about investment opportunities [66] or customers may have preferences about life insurance products [14]. Some research even frames stock trading use cases as a task with subjective outcomes, as some stocks may be preferred over others [25]. To add to this, while house prices may have a ground truth, in the case of finding a subtenant for an apartment, there is a tradeoff between the rental price and the certainty of finding a suitable candidate. This tradeoff also represents a subjective result, as some users may prefer to change a high rental rather than the certainty. Overall, tasks with subjective outcomes are most often found in the domains of personal preferences or ethical decision-making. In other domains, some articles frame a subjective task for users, where the AI system attempts to advise on the preferences of the user in opposition to an objective ground truth.

In addition to the use cases with subjective outcomes, the majority of literature is concerned with use cases that have an objective ground truth. Some of these cases are anchored in real-life scenarios, while others are isolated experiments. In the domain of *finance & business settings*, numerous articles are concerned with loan applications [38, 56, 62, 102] and stock trading [27, 28, 35]. Both appear to be favored use cases in current literature. One reason may be the close proximity to other machine learning research, which also frequently considers these cases. In addition to loan applications and stock trading, other use cases include basketball betting [43] and the prediction of incoming call center calls [13]. It is also noteworthy that in certain use cases (such as stock trading or incoming call prediction), the objective is to predict future events, such as the return of a stock or the number of incoming call center calls in the future. For these use cases with future events, decision-makers are confronted with delayed labels. Other use cases are concerned with directly uncovering the ground truth without the need for delayed labels. Closely adjacent to business decisions are the uses cases in *housing*, where users should mostly estimate the prices of houses [32, 33, 101] or find a fitting flat using online platforms [53]. Another domain that is often researched in AI reliance is the *medical* domain. The articles can be grouped into patient and elderly care [79, 133], image classification, such as MRI and X-ray [20, 46], and disease detection. Various diseases are considered, such as diabetes risk [40], myocardial infarction, a problem with the heart [98], oncology assessment [130], or sepsis classification [10]. In general, these articles also opted for experts, such as doctors, as decision-makers, as opposed to crowdworkers. This can be attributed to the high specialization required to detect diseases. The last groups of real-life anchored use cases are *autonomous vehicles* [3, 97, 120], *student performance* assessment [37, 105, 132], and the case of *recidivism* assessment, often based on the COMPAS algorithm [31, 50, 128]. In summary, we identify a multitude of use cases that investigate tasks in close proximity to real-life settings. We however also observe clear preferences in current research regarding AI reliance in domains such as finance and business, as well as in the medical domain.

For the isolated experiments, many articles ask the users for *logical decision-making*. As this is the most diverse category, we only present some examples. For instance, counting tasks are used, where people or objects in a picture have to be counted, and the AI provides advice on the number [15]. Some articles also consider games, such as chess [11], maze-solving [124], or custom path-building [127], where the AI presents advice on the correct strategy. Finally, some articles address what we refer to as general image, text, or video classification, which are tasks that do not fall within the scope of the other categories. Image classification tasks include noisy images based on the ImageNet dataset [81, 121], handwritten images from the MNIST dataset [26], bird images [21, 85, 112], or satellite images [21, 91]. The text classification tasks include sentiment analysis of reviews [8, 26, 76] or fake review detection [21, 114] and review credibility assessment [107], question answering tasks [8, 49], or translation tasks Mehandru et al. [88]. Finally, for video classification, video clips should be categorized [71] or activities should be recognized [96]. In addition to the real-world, anchored use cases, we observe that many articles are concerned with isolated experiments. Rather than providing insight into the specific domains in which AI systems are relied upon, these cases are employed to gain a general understanding of AI reliance.

In conclusion, a multitude of use cases have been identified in the current literature. Most of these tasks have objective correct advice. Some tasks also have subjective outcomes, where the AI aims to provide advice based on user preference. Additionally, it has been observed that some tasks are grounded in real-world scenarios (such as the financial or medical domain), while others are isolated experiments. Nevertheless, despite the multitude of real-world use cases, most articles collect data from online crowdworking platforms, such as MTurk and Prolific.

5.2 Measuring the Interaction of Human and AI

The second category of the categorization framework is the interaction between social and technical components. In the case of AI reliance, this interaction is human reliance, which we examine by reviewing current approaches and measures to quantify reliance on AI advice. First, we must establish a general understanding of the approach that should be used to measure reliance. As discussed below, there already seems to be some disagreement in the current literature regarding this matter. In light of the aforementioned approach, we proceed to present the existing measures of reliance.

Decision-Making Approach. Depending on whether we view AI systems as automation systems or as sources of advice, two distinct approaches can be identified. As described in Section 2, advice can be presented directly to the decision-maker or only shown after the decision-maker has already made an independent decision. This leads to two distinct approaches: a single-stage approach, where advice is presented before the decision is made, and a two-stage approach, where advice is presented after an initial decision is made and the decision-maker can revise their decision.

The more straightforward way to implement the decision-making approach is a *single-stage decision-making approach*, where users are directly confronted with the AI advice. The final decision from the human-AI interaction is recorded only then. This can also have some variants. Among the groups of articles that have a single-stage decision-making approach, there is one article that measures the human decision twice after being exposed to the AI decision [103], making it a multistage approach, but only for the human decision and not the AI advice. Further, instead of the human making their own decision after being exposed to the AI, sometimes the human can only accept or reject the AI advice [e.g. 14], or the human can decide if they want to decide, or if they want to let the AI decide for certain tasks [e.g. 97]. Another variant is that the AI advice is implemented automatically until a human intervention, as in the case of automatic driving games [e.g. 3]. While we find several variants, they are all unified as a single-stage approach, where the human is exposed to the AI advice without a prior decision on the task.
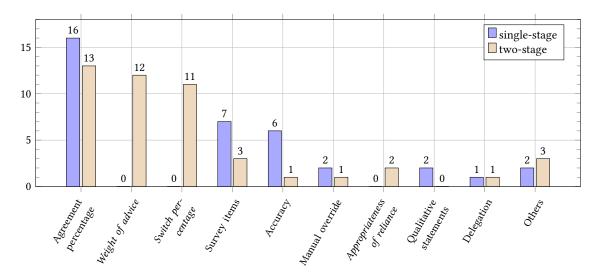
Fig. 5. Reliance measures, with some articles containing multiple measures. The *italics* measures strictly require a two-stage decision-making approach.

On the other hand, most articles employing a *two-stage* decision-making approach explicitly collect information about the initial human decision, for instance by allowing the human to make a decision for each task without the AI and allowing them to revise the decision in the face of AI advice. There are some exceptions to this pattern. One notable example is an article that measures preferences before and after being confronted with the AI system [94], as opposed to final decisions. Decision-making is not limited to two stages; rather, multistage approaches can exist. Morrison et al. [91]) employ such a multistage approach that presents more and more explanations to the users, and the user can adapt their decision after each step. Conversely, Elder et al. [43] do not explicitly collect an initial human decision, but introduce a waiting period before exposing the human to the AI advice, which also aims to enforce this two-stage approach. It is noteworthy that two-stage approaches have also been employed for training purposes, whereas the primary tasks did not use them. The users were instead permitted to delegate decisions to the AI system in the primary task [32]. With the exception of these few instances, all other articles employ a two-stage approach in accordance with expectations, explicitly collecting the initial human decision and allowing the human to revise after exposure to the AI advice.

Overall, we find a nearly even split between articles that have this two-stage approach (n = 36) and those with a single-stage approach, directly presenting AI advice without collecting the initial human decision (n = 35)[4]. The even split in the literature indicates that both approaches are perceived as applicable for research reliance on AI advice. However, both approaches entail different metrics that can be employed, as described in the following section.

Metrics and Measures. Upon examination of the existing literature, it becomes evident that a multitude of metrics are employed to assess the degree of reliance. Some of these metrics coincide with the above-introduced approaches to decision-making. In the single-stage decision-making approach, the number of agreements between human decisions and AI advice can be counted. In the two-stage approach, the number of switches from the initial human decision to

---

[4]One article explicitly tests both [46], which we count toward the set of articles with a two-stage approach

the AI advice can also be counted. Nevertheless, numerous additional metrics are employed in the literature, which are summarized in Figure 5.

The most commonly used measure of reliance is also one of the simplest, the *agreement percentage*. It expresses in how many decision instances a user has the same answer as the AI. It can be measured in both settings, the single-stage [e.g. 62, 122, 132] and the two-stage [e.g. 30, 56, 86]. It is therefore a convenient and intuitive measure, but some researchers argue that this match is not enough to measure reliance [115]. A downside might be that instances where the human simply agrees with the AI advice, but would have decided the same even without the AI advice are also counted toward this measure. Some articles therefore require the harder condition for reliance, that the human must change their decision because of the AI—leading to the necessity of the two-stage approach.

The *switch percentage* aims at this condition, as it measures the percentage of instances where the decision-maker switched their decision to the AI advice after being confronted with it. Interestingly, most articles that use switch percentage as a measure also use agreement percentage [e.g. 56, 85, 120]. Only a few articles exclusively use switch percentages [50, 94, 115]. Furthermore, Schmitt et al. [115] explicitly mentions the tension between agreement percentage and switch percentage in the current literature before opting exclusively for switch percentage. This shows that the agreement percentage is also seen as a viable measure. This two-stage approach has another advantage. If the use case is not a discrete decision case, but rather a regression case, the *WOA* (Section 2) can be used. Note that if the estimates and advice are binary, the WOA and switch percentage align. In other cases, this can give us an indication of the strength of the reliance, referring to how much the AI advice influenced the decision-maker's final decision. The idea of switching the decision was also used to define the *appropriateness of reliance* [112, 114] that counts the cases where reliance was appropriate and ignores overreliance and underreliance.

Several articles also use *survey items* to measure AI reliance. These ask participants to indicate on a (mostly 5- or 7-point) Likert scale how much they relied on the AI advice. Most of these articles exclusively use a survey to measure reliance [e.g. 35, 96, 102], while others also present other measures of reliance [24, 79]. In addition, two articles use a ranking of design features of the AI system's interface, such as explanations, where the feature with the highest ranking is said to be relied upon by users [26, 27]. Therefore, this measure is based on self-reports rather than observations of actual behavior. Another group of measures that use self-reports are *qualitative statements* from interviews about reliance [14, 101].

Another way reliance is sometimes measured is by using task *accuracy* as a proxy [e.g. 8, 21, 105]. In a within-subject design, a participant should make decisions either without AI advice or with AI advice. If performance of the final decisions improves, the participant is said to rely on AI advice. This is often studied in combination with the concept of complementary team performance, where the team consisting of a human and an AI should perform better than the individual members alone. Therefore, accuracy also indicates whether the user has appropriately applied reliance. Some articles are not interested in an active human decision but rather see reliance in the same way as a *non-action*, meaning a user should only intervene when the AI system tells them to [3, 43, 133]. If the user intervenes when the AI system does not tell them to, they have no reliance. Similarly, sometimes delegation is also measured as reliance [32, 97]. Finally, some other articles are concerned with reliance but employ tailored measures for that use case. For instance, in contrast to examining the result of the decision-making process, Kim et al. [71] assess reliance based on the frequency with which a user requests XAI assistance, as indicated by log files. Another behavioral metric that does not rely on implicit input decisions is the use of eye gaze as a measure of reliance Cao et al. [23]. Although this may be a viable option for measuring reliance, it is likely to be employed only in specific tasks to observe which aspects of the task the user is focusing on. Other articles attempt to identify patterns in user behavior to quantify reliance [20]

	Feedback on AI performance	No feedback on AI performance
Training with the AI	7	16
No training with AI	10	38

Table 5. Two-dimensional matrix on AI training and feedback.

or use regression analysis [46]. Interestingly, Radensky et al. [103] employ a single-stage approach that involves two human decisions following the AI advice, with a time interval between them. They use the discrepancy between the two human ratings as a measure of reliance, with higher discrepancies indicating reduced AI reliance. While these approaches offer insights into AI reliance, they are context-specific and may not be applicable to other studies.

Overall, there are evident clusters of measures that are employed with greater frequency. The most prevalent measure is the agreement and switch percentage. On occasion, surveys or qualitative statements are used as an additional indicator of reliance. This indicates that, first, there is no clear consensus on how to measure reliance, and second, reliance is a complex construct, and the use of a single measure is not always sufficient. We will examine the implications of various metrics being employed and the tensions between them in greater detail Section 6.

5.3 Human User as the Social Component

Once the settings and use cases have been established, as well as the general conceptualizations and measures of reliance, we proceed to the social component. In most decision-making tasks, both real-world and experimental, humans are confronted with multiple decision instances in rapid succession. For instance, a doctor may have numerous patients requiring a diagnosis. We claim that the experience gained from one instance may influence the decisions made in the next. We create a two-dimensional view and distinguish between training with the AI system and performance feedback during the main task, which is presented in Table 5. While analyzing the literature, it becomes apparent that not all articles explicitly state whether they provide feedback and/or training to the participants. Consequently, if no information about training with the AI was present, it was classified as no training. Furthermore, if an article tested out treatments with and without feedback or training, it was classified as an article with training or feedback present. Finally, if no information about feedback was present and the setting and use case did not provide feedback implicitly, it was classified as no feedback. An example of feedback provided implicitly is an autonomous vehicle driving game [3], where no crash with the autonomous vehicle indicates that the AI provided positive advice. Another example is a multiday trading game [28], where each day the return of the previous day was known.

As illustrated in Table 5, most articles do not provide training or feedback to users. We speculate that this could be due to underreporting. Another reason could be that a straighforward experiment design might lead to easier data collection than an experimental design with training and/ or feedback. Most articles do not examine changes in reliance over time or across multiple interactions, so that feedback would not have any impact on the empirical results. In certain instances, the provision of training and feedback may be impracticable, particularly in the case of tasks that are designed as one-off sessions [e.g. 14]. Some articles have training before the main task, where participants are trained on the task, for example bird image classification [112], but without the AI advice. One rationale for this approach is to assess the participants' skills before the main task, which could be a potential confounder to reliance. At the same time, the reliance behavior is not influenced by the training. In general, most articles employ a straightforward experimental design, where participants are not provided with any training with the AI system or performance feedback

on individual task instances. This approach is employed to measure the average reliance of the participants and to allow for comparisons between groups.

Another group of articles provide training to participants but no feedback. This is often done to reduce potential misunderstandings and to help users understand the function of the AI system. While some articles have training for the task without showing the AI advice to the user, this group focuses on articles where training with the AI advice was employed. Training involves allowing the user to interact with the system advice without considering those interactions in the analysis. It is sometimes done with one single decision instance [19], but often the training also involves many interactions with the AI system, such as 10 practice instances [32]. In some instances, the initial training is also integrated with an initial screening of participants, where performance in the training double as a metric for selecting potential participants [46]. In sum, we see several articles provide training with the AI system that does not count toward the final main task and metrics.

Some articles provide feedback but no training. The user can learn as they interact with the AI. This can also be seen as continuous training while on the task. However, all decision instances count toward the overall performance. While most articles in this group provide feedback after each decision, there are those that provide feedback halfway through the decision instances [37, 133]. Also, some articles do not explicitly state that feedback is provided. However, it can be inferred that users received feedback at least implicitly, so we also count them in this group. For example, in the case of music recommendations [103], a participant would directly know about the performance of the AI model based on the recommended music. These examples show that sometimes feedback is implicit and not necessarily a conscious design decision.

Finally, a few articles use both training and performance feedback during the main task. This is arguably a more complex experimental design, since not only do you have to make decisions about how and when to provide feedback, but you also have to design a training session for the users. This may be one reason why only a few articles choose this design. While this is rarely applied, for many applications it may be close to reality. A medical professional interacting with an AI system will most for example likely receive an introduction/training session with the tool and will also receive performance feedback, as they will be able to observe the patients and the system diagnosis. Most articles considering real-world use cases are therefore designed with training and feedback, such as the task of predicting the incoming calls of a business center [13], or the case of driving an autonomous car [3].

Overall, we find that most measures are calculated by having a user interact with the system multiple times and then taking the average of those interactions. For example, if a user interacts with the system 10 times and 8 times, the user gets the same output as the AI advice, the agreement percentage is 80%. However, this does not take into account the effects of repeated interactions with the AI system or prior exposure, such as training. A user who has been trained on the AI system may have a different reliance level than a user who is interacting with the system for the first time. Furthermore, if a user receives feedback on the performance of the AI system after each of these 10 decisions, the reliance might change due to this feedback. Only a few articles explicitly measure change in reliance. For example, Grgić-Hlača et al. [50] study the COMPAS algorithm and explicitly argue with real-world judges who also have the chance to adapt their reliance. Also, Leffrang et al. [80] investigates how users recover from bad advice given by the AI system. These cases are however rare, so most articles have neither training nor feedback for users.

5.4 AI System as the Technical Component

In addition to the social component, the technical component—the AI system—plays a pivotal role in sociotechnical systems. In this section, we therefore review the various AI systems used in the AI reliance literature, most prominently, the nature of the AI system itself and a common design decision of transparency mechanisms.

With regard to the AI system itself, it is evident that numerous articles do not construct real AI systems. Instead, users frequently interact with mockups and dummies in WOZ studies (n = 29) [e.g. 103, 115, 122]. In these studies, the users are informed that they are interacting with an AI system and receiving AI advice, whereas the researchers constructed this advice. This methodology is commonly employed, particularly in HCI research. It also streamlines data collection, as no AI system needs to be constructed. It is however crucial to exercise caution, as mocked systems do not provide real AI advice and may be biased in their output. Some of these studies even provide AI advice that is always correct Panigutti et al. [98], Srivastava et al. [120], therefore overreliance of users is not possible. As a preliminary step toward enabling decision-makers to interact with real AI systems, some articles construct real AI systems but present manually selected instances of AI advice. In other words, the researchers manually sample the dataset (n = 19) [e.g. 30, 62, 128]. This approach is frequently employed to achieve a specific performance of the model. One advantage of this approach is that it is closer to reality, as a real AI system created the advice. For instance, while some articles explicitly investigate outliers [e.g. 80], these are still hand-picked and expected, and the randomness is absent. These manually selected labels also distinguish between the model accuracy of the underlying model and the sampling accuracy of the samples presented to the decision-makers. For example, Chen et al. [30] presents eight decision instances, five of which have correct advice, resulting in a sampling accuracy of 62.5%, but the underlying models have accuracy above that. This distinction between sampling and model accuracy might induce biases. Only when decision-makers interact with real AI systems does sampling accuracy equal model accuracy. However, only a few articles allow users to interact with real AI systems (n = 21) [20, 102, 114]. In most cases this results in a random sample from a list of AI advice. Overall, we find that many studies do not use real AI systems but rather either fully mocked interfaces or hand-picked AI advice. To address the issue of WOZ systems and real AI systems, for example, Ashktorab et al. [5] tests both a real AI and a WOZ study with a "perfect" AI system, meaning a system that always outputs correct advice[5].

One of the most common design decision for AI systems are transparency mechanisms. For this review, we distinguish between explanations and the statement of AI performance or certainty. We find that many articles provide explanations to users (n = 27). There are numerous explanations. Notable examples include [14, 128], SHAP values [e.g. 38, 40]. Additionally, other forms of explanation have been employed, including those based on LIME [10] or counterfactual explanations [79]. Collectively, these examples illustrate the diversity of approaches to explanation. In addition, in some articles the user is provided a statement on the overall AI performance or the certainty on a specific decision task (n = 11) [e.g. 101, 103, 106]. Interestingly, only a few articles present both to the user (n = 8) [e.g. 21, 71, 105]. Finally, several articles do not employ any transparency mechanisms (n=25) [e.g. 23, 53, 86]. These articles often focus more on general applicability, such as the case of an AI that predicts preferences, where this premise is already the subject of research [e.g. 84, 122]. Other examples are articles that are interested in different aspects, such as the abovementioned change of reliance over time [50] or the use of multiple users [31]. It becomes apparent that a clear majority of articles present some form of transparency mechanism, which highlights their technology-centricity. Changes to the design of the technological system are often more interesting than the actual users, which often consist of crowdworkers and non-specific user groups.

---

[5]One article did not provide conclusive information about the implementation of the AI system and could not be categorized [25]

Overall, we find that several articles do not confront the users with real AI systems, but rather employ a WOZ methodology. Interestingly, we also find that only a few articles use real AI systems that the users interact with. We also find that many articles employ transparency mechanisms for their AI system, which might influence user reliance. The implications are discussed in the subsequent Section 6.

6 DISCUSSION

After a comprehensive review of the existing literature on AI reliance, we discuss its implications and shortcomings, which serve as inspiration for new research avenues. Overall, we identify the following five directions for future and more in-depth research on AI reliance: (1) the influence of various factors on reliance is not fully understood; (2) the external validity of studies is often limited; (3) different approaches have been employed to measure reliance, making comparisons difficult; (4) the impact of reliance on individuals and society over time has not been adequately addressed; (5) the potential implications of emerging issues related to AI reliance have not been fully explored. The remainder of this article will provide a more detailed examination of these avenues.

6.1 Limitations of Current AI Reliance Research

6.1.1 Unclear Influences on Reliance. The sociotechnical perspective on AI reliance identifies four distinct angles through which influences on AI reliance can be identified: the social component, the technical component, the interaction between these two, and the environment. All these factors play a role in AI reliance for decision-making. Only with this comprehensive perspective can we fully acknowledge the influences on AI reliance. A review of the literature reveals that most articles only consider individual components of the sociotechnical system. Even within these components, there is often no consensus on the factors that influence reliance on AI systems. In the following section, we discuss current approaches in the various components of sociotechnical systems. In conclusion, we urge further research to obtain a more comprehensive view of AI reliance.

A significant proportion of articles in this field focus on the technical component—the AI system itself—and its influence on reliance. The design aspects of AI systems are frequently the primary considerations, with explanations being a key area of interest. It is also evident in the literature that transparency mechanisms, such as explanations, do not have a monotonic influence, and explanations do not always lead to the same results [113]. Consequently, an exclusive focus on explanations, or the design of the technical component in general, is insufficient for a comprehensive understanding of AI reliance. It is imperative to consider the users interacting with the AI system.

The social component, referring to the human decision-maker, can also contribute to reliance. Humans are prone to cognitive biases [63] and AI systems have been shown to induce various such biases, most notably the anchoring or automation bias [105]. The effect of the bias may also depend on humans themselves, with some people being more prone to cognitive biases than others. For instance, domain knowledge has been shown to reduce these cognitive biases [72]. This may lead to higher self-reliance and, consequently, less reliance on AI systems. Some articles examine the effects of domain knowledge Bayer et al. [e.g. 11]. Consequently, the human itself, in terms of characteristics such as being experts or novices, may influence the level of reliance. This is considered in some articles, but there are additional considerations.

The environment in which humans and AI interact could also influence reliance. Some articles suggest that time pressure is an influence [23]. Barr Kumarakulasinghe et al. [10] point out that physicians' active work environment might also influence their reliance. To illustrate, a physician with the same computer vision system may exhibit excessive reliance on the system when a patient's condition is critical and a decision must be made in seconds. However, in an

environment where the physician has ample time to make decisions, there might be appropriate reliance on the AI system. The environment in which humans and AI interact may therefore influence the degree of AI reliance.

Finally, the interaction itself can influence reliance. A review of the literature reveals that there are two main approaches to decision-making and subsequently measuring reliance: single-stage and two-stage. Both influence the interaction and subsequently AI reliance. A single-stage decision-making approach facilitates fast decision-making and quick acceptance of the AI advice. In contrast, a two-stage approach enforces slow thinking, and the decision-maker must invest more effort in coming up with a decision. Cognitive forcing has been identified as a factor that influences reliance [19]. The interaction itself therefore plays a role in determining the extent to which humans rely on AI advice.

The influences on reliance can be conceptualized as a multidimensional space. For instance, a user interacting with the same AI system may exhibit varying degrees of reliance based on environmental factors, such as time constraints. Alternatively, a user interacting with the same AI system in the same environment may exhibit varying degrees of reliance based on the decision-making approach. This is because the same user may have a different reliance behavior when confronted with a single-stage decision-making process as opposed to a multistage approach. It is evident that current research articles tend to focus on a single (or at most two) dimension of AI reliance. A more comprehensive approach is however required to fully understand the nature of AI reliance. Specifically, there is a lack of understanding of how various dimensions or individual influence factors interact with each other with regard to reliance. For instance, it is likely that time pressure can interact with provision of explanations and their form. We therefore call for more research on all dimensions simultaneously. This would however lead to virtually unlimited combinations that would require attention. Nonetheless, we claim that focusing on combinations that occur frequently in the real world can help select appropriate combinations, while ensuring external validity.

6.1.2 Missing External Validity. The results show that reliance is considered in a multitude of use cases, ranging from autonomous vehicle control [e.g. 120] to house price predictions [e.g, 32] or solving ethical concerns [e.g. 122]. The results also show that these use cases are most often considered in controlled experiments, where WOZ approaches or manually picked sets of decision tasks dominate the interaction with the AI system. Further, most experiments are conducted with crowdworkers of MTurk or Prolific. While this has the positive effect of easy and fast data collection, it comes with the downside of missing external validity, as real AI systems with real users are rarely considered. The aspects and resulting consequences of this missing external validity are discussed in the following section.

The perception of randomness among humans is subject to bias. For instance, when asked to generate a random series, humans tend to generate series with higher-than-expected alternation rates. This is because humans tend to perceive clumps or streaks as not truly random, while they actually are Bar-Hillel and Wagenaar [9]. Conversely, AI is susceptible to randomness due to its probabilistic nature. Each time a study presents a manually selected set of decision instances or a WOZ study, users are confronted with a human bias from the outset. While numerous articles attempt to present a representative set of decision instances, they do not deliver any confirmation for the true randomness. Consequently, the users (decision-makers) of AI systems are rarely confronted with genuine and randomly selected AI output.

An illustrative example of a phenomenon that occurs in real-world AI interactions is the presence of outliers. They are an inherent and integral part of any AI system, and are often targeted to be detected [16]. While many approaches aim to minimize their occurrence, there is always the possibility of outliers occurring due to the probabilistic nature of current AI approaches. Examining the reliance on AI systems after a user is confronted with an outlier is therefore also

important, and is for example done by Leffrang et al. [80]. The users (decision-makers) of AI systems rarely encounter naturally occurring outliers in AI output in the current literature of AI reliance.

A further limitation of current research is that many studies use crowdworkers rather than actual end-users or decision-makers interacting with the system for its intended purpose. The use of crowdworkers, such as MTurk users, has been questioned and may render low data quality Chmielewski and Kucker [e.g. 34], Hsueh et al. [e.g. 60], Kennedy et al. [e.g. 68]. While a substantial body of research is dedicated to enhancing data quality and providing best practices Kees et al. [e.g. 67], concerns about the external validity and generalizability of the results persist. A recent discussion has emerged regarding the potential for crowdworkers to employ AI systems, such as LLMs, to automate their tasks [64]. This would also diminish the reliability of using crowdworkers, as the responses are not human-generated but rather AI-generated, removing the human element from the equation. This would result in investigations of the reliance of AI systems on other AI systems.

It becomes evident that real users and real environments are rarely studied. This also raises the question of high-stakes decision-making tasks. While several of the examined use cases may be considered high-stakes decisions, such as medical decision-making [e.g. 40, 98, 130], the consequences of these decisions are rarely observed. For instance, in the case of kidney donors [94], no actual patient is affected by the decision, therefore the decision-maker may approach the problem differently if they were aware that a real patient's life or death depended on the outcome. Consequently, we observe a lack of high-stakes decisions in the current literature on AI reliance.

Overall, the current research lacks sufficient external validity, suggesting that the decision-making scenarios in the studies may not accurately reflect real-world decision-making. We therefore call for research that examines more real-world settings and existing AI systems. There are numerous existing AI systems that could be used to investigate reliance. For instance, researchers with access to Netflix or Amazon data could investigate user reliance when faced with their recommendation algorithms. While there is undoubtedly a considerable amount of research on these algorithms [4, 54, 83], it seems to be concerned with the accuracy and design of the AI systems themselves rather than user reliance. While reliance is related to accuracy, it is not the same. To illustrate, if the Netflix algorithm were to predict a movie recommendation with 100% accuracy based on the training data, this performance could only be achieved in real life if the user relies on these recommendations. This reliance has more facets than the pure performance of the AI advice, which is highlighted by the sociotechnical perspective. Consequently, it is not only the performance that must be evaluated but also the extent of AI reliance, particularly in real-world settings.

6.1.3 Conflicting Approaches to Measuring Reliance. One potential explanation for the lack of external validity in the current literature may be the difficulty of measuring reliance. In particular, in real-world settings, it is unclear whether a human decided to do something because of the AI or whether the human simply agreed with the AI, but the AI did not influence their decision-making. For example, if a consumer purchases a product after it has been recommended by Amazon's algorithm, it is challenging to ascertain whether the consumer would have made the purchase regardless of the recommendation. Conversely, the question of whether a human decision-maker was influenced by AI advice becomes more straightforward in experiments where researchers can control the environment.

In controlled experiments, research sometimes employs a two-stage decision-making approach where the user must first make their own decision without any AI support. They get AI advice only afterwards and use it to change their decision. A change can then be attributed to the AI advice and be defined as reliance. In contrast, the single-stage decision approach directly confronts the user with the AI advice, rendering any such consideration impossible.

This two-stage approach offers several advantages. It is designed to counteract some human cognitive biases, such as anchoring or priming, which occur by presenting the AI advice first. Additionally, research in the field of advice-taking indicates that this approach can lead to enhanced performance overall Sniezek and Buckley [118]. The two-step approach allows for the identification of instances where a user may have agreed with the AI advice but would have reached the same conclusion independently. This approach also has a potential drawback, as humans may be reluctant to alter their initial decisions. One indication of status quo bias is that humans are cautious about changing the status quo [109]. This is evidenced by their inclination to remain at the current status, even when presented with new information. Furthermore, the initial answer could be seen as self-anchoring, with the individual's initial response becoming a reference point for subsequent decisions.
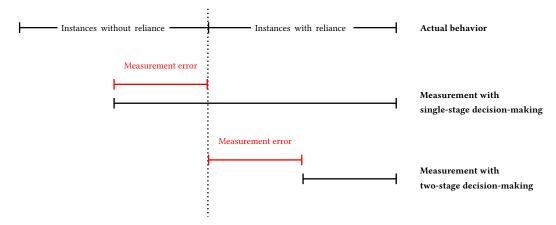

Fig. 6. Tradeoff between metrics for single-stage and two-stage decision-making processes. Note: Boxes are not true to size and are for illustration purposes only

Both approaches use a multitude of metrics, as illustrated in Figure 5 on page 17. While some of these metrics are directly inferred from either approach, such as agreement percentage and switch percentage, others are only loosely connected, such as survey items and qualitative statements. In particular, the latter metrics use self-reported data and are therefore suboptimal for collecting behavioral data. Consequently, we view them as supplementary rather than standalone metrics. Concerning the single-stage and two-stage approaches, analogies can be drawn between the concepts of precision and recall. While the single-stage approach increases recall of instances where the human relies on the AI system, the two-stage approach increases precision. These potential errors are illustrated in Figure 6. It is currently not possible to quantify these measurement errors, consequently it is not possible to determine which approach is superior. The choice between the two approaches depends on the specific errors that one wishes to avoid. The decision to select either one must therefore be made with careful consideration.

In summary, we see that both the single-stage and two-stage approaches have advantages and disadvantages. The single-stage approach is easy to measure, straightforward, and understandable. Reliance is however overestimated due to the classification of numerous instances as such, despite the user ultimately being the decision-maker. The two-stage approach is more robust, therefore reliance can be identified with greater precision. Instances where a user makes a different decision because they have to make a prior decision may however be overlooked, therefore it is important to make conscious decisions about which approach to use. More research should be conducted to investigate the differences between the two approaches.
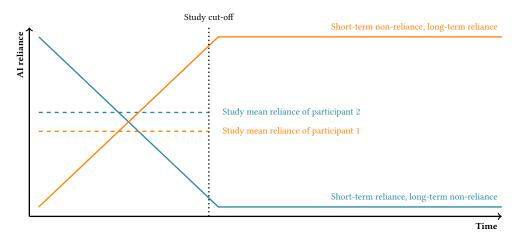
Fig. 7. Possible reliance change over time where results unaware of the change over time might lead to misleading conclusions.

6.1.4   Disregard for Reliance Changes Over Time. Most articles investigate users interacting with an AI system on multiple occasions. The reliance level (e.g. agreement fraction) is then calculated by averaging all decision instances. For instance, if a user relies on the AI system eight out of ten times, the reliance level is 0.8. Alternatively, if the WOA is calculated, the reliance is defined as the mean WOA over these ten decision instances. This average reliance consideration neglects potential trends of reliance behavior. When a user interacts with an AI system on multiple occasions, it is possible that the effect may only become apparent over time. However, only a few articles consider the changes to reliance over multiple interactions, for exampleLeffrang et al. [80] focused on the effect of an outlier and how reliance changes after this outlier.

To illustrate this point, consider a study that investigates reliance on AI advice and concludes that users, in general, have a moderate level of reliance. A longitudinal field trial may however reveal that after some time some users start to always rely on AI advice, while others stop. Consequently, the findings of the study may become invalid after a certain period of time. An extreme case of this phenomenon is illustrated in Figure 7. In Figure 7 the (hypothetical) participant 1 (orange) increases the reliance over the time of the study and over time will always rely on the AI advice, which we denote as *short-term non-reliance, long-term reliance*. The (hypothetical) participant 2 (blue) in Figure 7 is the opposite, which we denote as *short-term reliance, long-term non-reliance*. While both will have a similar reliance behavior in a study, both users have a fundamental different reliance over a longer period of time. A study that only measures average reliance will not detect this decline. This also has implications for the real-life use of AI systems. For example, if a new AI system is tested and the study concludes that reliance is high, in practice the same tool might not impact the real world if it follows a similar pattern as (hypothetical) participant 2 in Figure 7. Consequently, we posit that studies should be guided by a conscious decision as to whether to consider the reliance change over time or to limit their scope to the average reliance level.

Overall, it is important to consider the concept of reliance over time. Most articles do not require any changes to the experimental design; rather, they should simply report on any changes in reliance. Nevertheless, there is a clear need for further research into this phenomenon of reliance on the part of users changing over time. This will help to identify the antecedents and consequences of this process, which can then be addressed in scientific literature.

6.2 Emerging Issues on AI Reliance

After reviewing and analyzing the existing literature, we come up with several emerging issues that we believe require further investigation. First, there is the emerging issue of reliance on generative AI systems. These systems possess the capability of providing a new class of AI output, such as creative texts. We believe that new conceptualizations and measures are required for this class of problem. Second, we observe a focus on settings with one human decision-maker in the current literature. Real-world settings frequently comprise multiple humans and decision-makers. There is a pressing need to investigate this emerging issue.

Reliance on Generative AI Output. In recent years, the topic of generative AI has emerged as a significant area of interest. These systems possess capabilities that do not fit directly into any of the discourses presented in Section 2. The crucial difference between generative AI and more traditional AI systems is that generative AI models can create output where humans might not be able to make an initial decision. In the context of image generation models, it is for instance possible that the human user may have a general idea of the outcome, but it is unlikely that they would be able to form an initial decision on a creative outcome. Consequently, the forthcoming generative AI systems may form an additional discourse to AI reliance research.

It is noteworthy that some of the literature on AI reliance already employs generative AI models, most notably large language models, as evidenced by the work of Schmitt et al. [115] and Goyal et al. [49]. These approaches use large language models as the underlying foundation, yet the output remains a closed form of advice, such as an answer to a specific question, as opposed to a creative and generative task.

The most prevalent example of generative advice currently is GPT models, which are accessible through the interface of ChatGPT [1]. The advent of generative AI systems has enabled everyday users to use these systems, but it is not yet clear how they are integrated into users' decision-making process. For instance, if a system preformulates a response to an email and the human user adapts some passages, it is unclear what reliance really means in that case. In the context of text generation, the text similarity between the LLM-generated output and the human text based on this might be used as a quantification. This also raises the question of text similarity, as computer science has developed numerous metrics at different levels of granularity, including character-level, word-level, and even semantic similarity.

As these systems are increasingly used, there is a pressing need to conceptualize and define reliance on these systems. Once this conceptualization is established, research should aim to identify methods for quantifying reliance on generative AI output. Future research should investigate how current measures of reliance, such as agreement fraction or weight of advice, might be applied to generative AI systems. The need for new measures arises with the increasing use of generative AI systems.

Reliance with more than one user. In the context of research on human-AI teams, most studies focus on a single user and AI system. Similarly, research on the performance of human-AI teams tends to examine the combined output of a single user and AI system, with the expectation that this team will outperform the individual components. Given this focus on single users and AI systems, it is not surprising that the corpus of relevant AI reliance articles predominantly discusses the performance of a single user. There is only one article that considers the performance of multiple users. Chiang et al. [31] examine group decision-making in the context of the COMPAS algorithm.

In real-world settings, we posit that most cases involve multiple users or even multiple parties with conflicting interests. For instance, numerous articles address house price predictions with a focus on a single user [32, 33, 101]. The impact of such AI systems on real-life settings is contingent upon the degree of reliance placed upon them by both

Category	Attribute				
Environment	Task	Objective tasks		Subjective tasks	
	Setting	Crowdworkers / Experts		Students / Other	
	Use case	...			
Interaction	Decision-making approach	Single-stage		Two-stage	
	Reliance measure	Agreement Percentage, Weight of Advice, Switch Percentage, Survey Items, Accuracy		Manual Override, Appropriateness of Reliance, Qualitative Statements, Delegation, Other	
Social	User training	None	Training without AI advice	Training with AI advice	Other
	Performance feedback	None	Partial feedback	Full feedback	Other
Technical	AI system training	WOZ	Manual samples	Random samples	Other
	Transparency mechanism	None	Performance metrics	Explanations	Other

Fig. 8. Morphological box for the design of AI reliance studies.

the seller and the buyer of houses. Given the inherent conflict of interest between these two parties, further research is warranted to investigate the extent to which these systems can be relied upon.

In addition to the existing literature on AI reliance, recent studies investigate the integration of AI in creative, collaborative settings [117]. The objective of the AI system is to support the ideation of multiple users. These settings are however all concerned with collaborative settings, where multiple users come together with common goals and interests, excluding settings with conflicting interests.

In conclusion, it is evident that there is a necessity to consider settings with multiple users and parties that are able to capture real human interaction. This can be achieved by extending the currently considered use cases to encompass more stakeholders. As previously discussed, AI systems supporting house price predictions must be relied upon by both the buyer and seller. Similarly, AI systems supporting doctors in diagnoses must also be relied upon by the patients in order to have real-life impact.

6.3 A Morphological Box for Sound AI Reliance Research

A review of the extant literature on AI reliance reveals that the articles exhibit a number of common patterns, which can be grouped according to the concepts presented in Table 4 on page 12. It also became evident that the requisite information could at times not be extracted with ease, as some articles lacked clarity. As previously indicated, user training was a notable example. We invite all studies on AI reliance to use the concepts introduced in Table 4 on page 12 as a basis for presenting the required information. To facilitate the work of researchers, we have transformed this table into a morphological box. This morphological box is presented in Figure 8. The attributes for the morphological box are based on the findings concerning general patterns in the literature and presentations in the results section. The

morphological box can serve as guidance for all future studies on AI reliance. It is not our intention to assert that some approaches or attributes are inherently superior; this morphological box should however serve as a reference for the design of studies. It is also recommended that researchers provide a rationale for the selection of specific approaches. For instance, future studies on AI reliance should clearly justify the choice of a single-stage over a two-stage approach or the decision to have user training. By doing so, we can establish a unified framework for AI reliance studies and provide guidance to researchers in the design of their studies. Ultimately, this leads to a common understanding and comparable results of AI reliance researchers.

In addition to using this morphological box as guidance, research can also be conducted with the objective of expanding it. This review represents an initial and indispensable step toward a unified and guiding framework for AI reliance researchers. The morphological box in reference Figure 8 just reflects the current state of research. Future research may present extensions to it, either by introducing new attributes for the individual subconcepts or by introducing new subconcepts altogether. It is probable that research on the emerging issues presented in Section 6.2 will result in the identification of new subconcepts. These may for instance include the underlying interaction structure between user(s) and AI, should more than one user be considered. Before extending the morphological box, further substantial research on the emerging issues is recommended.

7 CONCLUSION

In this study, we conduct a survey of existing literature on AI reliance. We employ a sociotechnical perspective, as this allows us to gain a comprehensive understanding. We derive concepts based on the four components of a sociotechnical system to classify AI reliance literature and to classify current literature. Furthermore, we discuss current issues and topics related to AI reliance.

This review aims to support future researchers on AI reliance. We provide researchers with a framework to evaluate their AI systems, enabling them to classify and present their results. We also identify future avenues for research on AI reliance, assisting researchers in their endeavors.

REFERENCES

[1] 2022. Introducing ChatGPT. https://openai.com/blog/chatgpt
[2] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052
[3] Ighoyota Ben Ajenaghughrure, Sonia Cláudia Da Costa Sousa, and David Lamas. 2021. Psychophysiological Modeling of Trust In Technology: Influence of Feature Selection Methods. *Proceedings of the ACM on Human-Computer Interaction* 5, EICS (May 2021), 1–25. https://doi.org/10.1145/3459745
[4] Xavier Amatriain and Justin Basilico. 2015. Recommender systems in industry: A netflix case study. In *Recommender systems handbook*. Springer, 385–419.
[5] Zahra Ashktorab, Michael Desmond, Josh Andres, Michael Muller, Narendra Nath Joshi, Michelle Brachman, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Christine T. Wolf, Evelyn Duesterwald, Casey Dugan, Werner Geyer, and Darrell Reimer. 2021. AI-Assisted Human Labeling: Batching for Efficiency without Overreliance. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 1–27. https://doi.org/10.1145/3449163
[6] Annette Baier. 1986. Trust and Antitrust. *Ethics* 96 (1986), 231–260.
[7] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–16.
[8] 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*. ACM, Yokohama Japan, 1–16. https://doi.org/10.1145/3411764.3445717
[9] Maya Bar-Hillel and Willem A Wagenaar. 1991. The perception of randomness. *Advances in applied mathematics* 12, 4 (1991), 428–454.

[10] Nesaretnam Barr Kumarakulasinghe, Tobias Blomberg, Jintai Liu, Alexandra Saraiva Leao, and Panagiotis Papapetrou. 2020. Evaluating Local Interpretable Model-Agnostic Explanations on Clinical Machine Learning Classification Models. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, Rochester, MN, USA, 7–12. https://doi.org/10.1109/CBMS49503.2020.00009

[11] Sarah Bayer, Henner Gimpel, and Moritz Markgraf. 2022. The role of domain expertise in trusting and following explainable AI decision support systems. *Journal of Decision Systems* 32, 1 (Dec. 2022), 110–138. https://doi.org/10.1080/12460125.2021.1958505

[12] Natalie C Benda, Laurie L Novak, Carrie Reale, and Jessica S Ancker. 2022. Trust in AI: why we should be designing for APPROPRIATE reliance. *Journal of the American Medical Informatics Association* 29, 1 (2022), 207–212.

[13] Benedikt Berger, Martin Adam, Alexander Rühr, and Alexander Benlian. 2021. Watch Me Improve—Algorithm Aversion and Demonstrating the Ability to Learn. *Business & Information Systems Engineering* 63, 1 (Feb. 2021), 55–68. https://doi.org/10.1007/s12599-020-00678-5

[14] Astrid Bertrand, James R. Eagan, and Winston Maxwell. 2023. Questioning the ability of feature-based explanations to empower non-experts in robo-advised financial decision-making. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 943–958. https://doi.org/10.1145/3593013.3594053

[15] Eric Bogert, Aaron Schecter, and Richard T. Watson. 2021. Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific Reports* 11, 1 (April 2021), 8028. https://doi.org/10.1038/s41598-021-87480-9

[16] Azzedine Boukerche, Lining Zheng, and Omar Alfandi. 2020. Outlier detection: Methods, models, and classification. *ACM Computing Surveys (CSUR)* 53, 3 (2020), 1–37.

[17] 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. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–27. https://doi.org/10.1145/3555212

[18] Tim Brennan and William Dieterich. 2018. Correctional offender management profiles for alternative sanctions (COMPAS). *Handbook of recidivism risk/needs assessment tools* (2018), 49–75.

[19] 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. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 1–21. https://doi.org/10.1145/3449287 arXiv: 2102.09692.

[20] 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*. ACM, Hamburg Germany, 1–20. https://doi.org/10.1145/3544548.3581095

[21] Ángel Alexander Cabrera, Adam Perer, and Jason I. Hong. 2023. Improving Human-AI Collaboration With Descriptions of AI Behavior. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 1–21. https://doi.org/10.1145/3579612

[22] Deborah Ann Cafarelli. 1998. *Effect of false alarm rate on pilot use and trust of automation under conditions of simulated high risk.* Ph. D. Dissertation. Massachusetts Institute of Technology.

[23] Shiye Cao, Catalina Gomez, and Chien-Ming Huang. 2023. How Time Pressure in Different Phases of Decision-Making Influences Human-AI Collaboration. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Sept. 2023), 1–26. https://doi.org/10.1145/3610068

[24] Shiye Cao and Chien-Ming Huang. 2022. Understanding User Reliance on AI in Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–23. https://doi.org/10.1145/3555572

[25] Noah Castelo, Maarten W. Bos, and Donald R. Lehmann. 2019. Task-Dependent Algorithm Aversion. *Journal of Marketing Research* 56, 5 (Oct. 2019), 809–825. https://doi.org/10.1177/0022243719851788

[26] 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 Transactions on Interactive Intelligent Systems* (March 2023), 3588320. https://doi.org/10.1145/3588320

[27] 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*. ACM, Sydney NSW Australia, 251–263. https://doi.org/10.1145/3581641.3584080

[28] Alvaro Chacon, Edgar E. Kausel, and Tomas Reyes. 2022. A longitudinal approach for understanding algorithm use. *Journal of Behavioral Decision Making* 35, 4 (Oct. 2022), e2275. https://doi.org/10.1002/bdm.2275

[29] Sutirtha Chatterjee, Suprateek Sarker, Michael J Lee, Xiao Xiao, and Amany Elbanna. 2021. A possible conceptualization of the information systems (IS) artifact: A general systems theory perspective 1. *Information Systems Journal* 31, 4 (2021), 550–578.

[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. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Sept. 2023), 1–32. https://doi.org/10.1145/3610219

[31] 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*. ACM, Hamburg Germany, 1–18. https://doi.org/10.1145/3544548.3581015

[32] C.-W. Chiang and M. Yin. 2021. You'd Better Stop! Understanding Human Reliance on Machine Learning Models under Covariate Shift. In *ACM Int. Conf. Proc. Ser.* Association for Computing Machinery, 120–129. https://doi.org/10.1145/3447535.3462487 Journal Abbreviation: ACM Int. Conf. Proc. Ser..

[33] C.-W. Chiang and M. Yin. 2022. Exploring the Effects of Machine Learning Literacy Interventions on Laypeople's Reliance on Machine Learning Models. In *Int Conf Intell User Interfaces Proc IUI.* Association for Computing Machinery, 148–161. https://doi.org/10.1145/3490099.3511121 Journal Abbreviation: Int Conf Intell User Interfaces Proc IUI.

[34] Michael Chmielewski and Sarah C Kucker. 2020. An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science* 11, 4 (2020), 464–473.

[35] Alton Y.K. Chua, Anjan Pal, and Snehasish Banerjee. 2023. AI-enabled investment advice: Will users buy it? *Computers in Human Behavior* 138 (Jan. 2023), 107481. https://doi.org/10.1016/j.chb.2022.107481

[36] Trevor Deley and Elizabeth Dubois. 2020. Assessing trust versus reliance for technology platforms by systematic literature review. *Social Media+ Society* 6, 2 (2020), 2056305120913883.

[37] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2018. Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science* 64, 3 (March 2018), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

[38] M. Dikmen and C. 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). https://doi.org/10.1016/j.ijhcs.2022.102792 Publisher: Academic Press.

[39] Mateusz Dolata, Stefan Feuerriegel, and Gerhard Schwabe. 2022. A sociotechnical view of algorithmic fairness. *Information Systems Journal* 32, 4 (2022), 754–818.

[40] Yuhan Du, Anna Markella Antoniadi, Catherine McNestry, Fionnuala M. McAuliffe, and Catherine Mooney. 2022. The Role of XAI in Advice-Taking from a Clinical Decision Support System: A Comparative User Study of Feature Contribution-Based and Example-Based Explanations. *Applied Sciences* 12, 20 (Oct. 2022), 10323. https://doi.org/10.3390/app122010323

[41] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, and Rajiv Ranjan. 2023. Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Comput. Surv.* 55, 9, Article 194 (jan 2023), 33 pages. https://doi.org/10.1145/3561048

[42] 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.

[43] Hannah Elder, Casey Canfield, Daniel B. Shank, Tobias Rieger, and Casey Hines. 2022. Knowing When to Pass: The Effect of AI Reliability in Risky Decision Contexts. *Human Factors: The Journal of the Human Factors and Ergonomics Society* (May 2022), 001872082211006. https://doi.org/10.1177/00187208221100691

[44] EU. 2021. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[45] EU. 2024. European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=EP%3AP9_TA%282024%290138

[46] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. 2022. Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging. In *2022 ACM Conference on Fairness, Accountability, and Transparency.* ACM, Seoul Republic of Korea, 1362–1374. https://doi.org/10.1145/3531146.3533193

[47] Krzysztof Z Gajos and Lena Mamykina. 2022. Do people engage cognitively with AI? Impact of AI assistance on incidental learning. In *27th international conference on intelligent user interfaces.* 794–806.

[48] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning.* MIT press.

[49] Navita Goyal, Eleftheria Briakou, Amanda Liu, Connor Baumler, Claire Bonial, Jeffrey Micher, Clare Voss, Marine Carpuat, and Hal Daumé III. 2023. What Else Do I Need to Know? The Effect of Background Information on Users' Reliance on QA Systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 3313–3330. https://doi.org/10.18653/v1/2023.emnlp-main.201

[50] Nina Grgić-Hlača, Christoph Engel, and Krishna P. Gummadi. 2019. Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–25. https://doi.org/10.1145/3359280

[51] Ross Gruetzemacher and David Paradice. 2022. Deep transfer learning & beyond: Transformer language models in information systems research. *ACM Computing Surveys (CSUR)* 54, 10s (2022), 1–35.

[52] Ziyang Guo, Yifan Wu, Jason D Hartline, and Jessica Hullman. 2024. A Decision Theoretic Framework for Measuring AI Reliance. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency.* 221–236.

[53] Akshit Gupta, Debadeep Basu, Ramya Ghantasala, Sihang Qiu, and Ujwal Gadiraju. 2022. To Trust or Not To Trust: How a Conversational Interface Affects Trust in a Decision Support System. In *Proceedings of the ACM Web Conference 2022.* ACM, Virtual Event, Lyon France, 3531–3540. https://doi.org/10.1145/3485447.3512248

[54] Blake Hallinan and Ted Striphas. 2016. Recommended for you: The Netflix Prize and the production of algorithmic culture. *New media & society* 18, 1 (2016), 117–137.

[55] Katherine Hawley. 2014. Trust, distrust and commitment. *Noûs* 48, 1 (2014), 1–20.

[56] 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. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Sept. 2023), 1–29. https://doi.org/10.1145/3610067

[57] 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*. ACM, Hamburg Germany, 1–18. https://doi.org/10.1145/3544548.3581025

[58] Richard Holton. 1994. Deciding to trust, coming to believe. *Australasian journal of philosophy* 72, 1 (1994), 63–76.

[59] Yoyo Tsung-Yu Hou and Malte F. Jung. 2021. Who is the Expert? Reconciling Algorithm Aversion and Algorithm Appreciation in AI-Supported Decision Making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–25. https://doi.org/10.1145/3479864

[60] Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, Eric Ringger, Robbie Haertel, and Katrin Tomanek (Eds.). Association for Computational Linguistics, Boulder, Colorado, 27–35. https://aclanthology.org/W09-1904

[61] Mandy Hütter and Klaus Fiedler. 2019. Advice taking under uncertainty: The impact of genuine advice versus arbitrary anchors on judgment. *Journal of Experimental Social Psychology* 85 (Nov. 2019), 103829. https://doi.org/10.1016/j.jesp.2019.103829

[62] Johannes Jakubik, Jakob Schöffer, Vincent Hoge, Michael Vössing, and Niklas Kühl. 2022. An empirical evaluation of predicted outcomes as explanations in human-AI decision-making. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 353–368.

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

[64] Krystal Kauffman and Adrienne Williams. 2023. Turk Wars: How AI Threatens the Workers Who Fuel It. (2023). https://doi.org/10.48558/NRZX-6Q03 Publisher: Stanford Social Innovation Review.

[65] Davinder Kaur, Suleyman Uslu, Kaley J Rittichier, and Arjan Durresi. 2022. Trustworthy artificial intelligence: a review. *ACM computing surveys (CSUR)* 55, 2 (2022), 1–38.

[66] Christoph Keding and Philip Meissner. 2021. Managerial overreliance on AI-augmented decision-making processes: How the use of AI-based advisory systems shapes choice behavior in R&D investment decisions. *Technological Forecasting and Social Change* 171 (Oct. 2021), 120970. https://doi.org/10.1016/j.techfore.2021.120970

[67] Jeremy Kees, Christopher Berry, Scot Burton, and Kim Sheehan. 2017. An analysis of data quality: Professional panels, student subject pools, and Amazon's Mechanical Turk. *Journal of advertising* 46, 1 (2017), 141–155.

[68] Ryan Kennedy, Scott Clifford, Tyler Burleigh, Philip D Waggoner, Ryan Jewell, and Nicholas JG Winter. 2020. The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods* 8, 4 (2020), 614–629.

[69] Charalampia Xaroula Kerasidou, Angeliki Kerasidou, Monika Buscher, and Stephen Wilkinson. 2022. Before and beyond trust: reliance in medical AI. *Journal of medical ethics* 48, 11 (2022), 852–856.

[70] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2022. Transformers in vision: A survey. *ACM computing surveys (CSUR)* 54, 10s (2022), 1–41.

[71] Chris Kim, Xiao Lin, Christopher Collins, Graham W. Taylor, and Mohamed R. Amer. 2021. Learn, Generate, Rank, Explain: A Case Study of Visual Explanation by Generative Machine Learning. *ACM Transactions on Interactive Intelligent Systems* 11, 3-4 (Dec. 2021), 1–34. https://doi.org/10.1145/3465407

[72] Josef F Krems and Christoph Zierer. 1994. Are experts immune to cognitive bias? Dependence of" confirmation bias" on specialist knowledge. *Zeitschrift fur experimentelle und angewandte Psychologie* 41, 1 (1994), 98–115.

[73] Olya Kudina and Ibo van de Poel. 2024. A sociotechnical system perspective on AI. *Minds and Machines* 34, 3 (2024), 21.

[74] Niklas Kühl, Max Schemmer, Marc Goutier, and Gerhard Satzger. 2022. Artificial intelligence and machine learning. *Electronic Markets* 32, 4 (2022), 2235–2244.

[75] 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 (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1369–1385. https://doi.org/10.1145/3593013.3594087

[76] Vivian Lai, Yiming Zhang, Chacha Chen, Q. Vera Liao, and Chenhao Tan. 2023. Selective Explanations: Leveraging Human Input to Align Explainable AI. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Sept. 2023), 1–35. https://doi.org/10.1145/3610206

[77] Allen S Lee. 2004. Thinking about social theory and philosophy for information systems. *Social theory and philosophy for information systems* 1 (2004), 26.

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

[79] 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. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Sept. 2023), 1–22. https://doi.org/10.1145/3610218

[80] Dirk Leffrang, Kevin Bösch, and Oliver Müller. 2023. Do People Recover from Algorithm Aversion? An Experimental Study of Algorithm Aversion over Time. https://hdl.handle.net/10125/103122

[81] Heliodoro Tejeda Lemus, Aakriti Kumar, and Mark Steyvers. 2022. An Empirical Investigation of Reliance on AI-Assistance in a Noisy-Image Classification Task. (2022).

[82] Paul M Leonardi. 2012. Materiality, sociomateriality, and socio-technical systems: What do these terms mean? How are they different? Do we need them. *Materiality and organizing: Social interaction in a technological world* 25, 10 (2012), 1093.

[83] Greg Linden, Brent Smith, and Jeremy York. 2003. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing* 7, 1 (2003), 76–80.

[84] 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 (March 2019), 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

[85] Zhuoran Lu, Zhuoyan Li, Chun-Wei Chiang, and Ming Yin. 2023. Strategic Adversarial Attacks in AI-assisted Decision Making to Reduce Human Trust and Reliance. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Macau, SAR China, 3020–3028. https://doi.org/10.24963/ijcai.2023/337

[86] Z. Lu and M. Yin. 2021. Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *Conf Hum Fact Comput Syst Proc*. Association for Computing Machinery. https://doi.org/10.1145/3411764.3445562 Journal Abbreviation: Conf Hum Fact Comput Syst Proc.

[87] John McCarthy, Marvin L Minsky, Nathaniel Rochester, and Claude E Shannon. 2006. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine* 27, 4 (2006), 12–12.

[88] Nikita Mehandru, Sweta Agrawal, Yimin Xiao, Ge Gao, Elaine Khoong, Marine Carpuat, and Niloufar Salehi. 2023. Physician Detection of Clinical Harm in Machine Translation: Quality Estimation Aids in Reliance and Backtranslation Identifies Critical Errors. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 11633–11647. https://doi.org/10.18653/v1/2023.emnlp-main.712

[89] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* 54, 6 (2021), 1–35.

[90] David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G Altman, and PRISMA Group*. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine* 151, 4 (2009), 264–269.

[91] Katelyn Morrison, Donghoon Shin, Kenneth Holstein, and Adam Perer. 2023. Evaluating the Impact of Human Explanation Strategies on Human-AI Visual Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 1–37. https://doi.org/10.1145/3579481

[92] Bonnie M Muir. 1987. Trust between humans and machines, and the design of decision aids. *International journal of man-machine studies* 27, 5-6 (1987), 527–539.

[93] Bonnie M Muir. 1994. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics* 37, 11 (1994), 1905–1922.

[94] Saumik Narayanan, Guanghui Yu, Chien-Ju Ho, and Ming Yin. 2023. How does Value Similarity affect Human Reliance in AI-Assisted Ethical Decision Making?. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Montr\'{e}al QC Canada, 49–57. https://doi.org/10.1145/3600211.3604709

[95] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 72–78.

[96] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *26th International Conference on Intelligent User Interfaces*. ACM, College Station TX USA, 340–350. https://doi.org/10.1145/3397481.3450639

[97] Kazuo Okamura and Seiji Yamada. 2020. Adaptive trust calibration for human-AI collaboration. *PLOS ONE* 15, 2 (Feb. 2020), e0229132. https://doi.org/10.1371/journal.pone.0229132

[98] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. 2022. Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–9. https://doi.org/10.1145/3491102.3502104

[99] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[100] Dana Pessach and Erez Shmueli. 2022. A Review on Fairness in Machine Learning. 55, 3, Article 51 (feb 2022), 44 pages. https://doi.org/10.1145/3494672

[101] 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*. ACM, Sydney NSW Australia, 379–396. https://doi.org/10.1145/3581641.3584033

[102] Erasmo Purificato, Flavio Lorenzo, Francesca Fallucchi, and Ernesto William De Luca. 2023. The Use of Responsible Artificial Intelligence Techniques in the Context of Loan Approval Processes. *International Journal of Human–Computer Interaction* 39, 7 (April 2023), 1543–1562. https://doi.org/10.1080/10447318.2022.2081284

[103] Marissa Radensky, Julie Anne Séguin, Jang Soo Lim, Kristen Olson, and Robert Geiger. 2023. "I Think You Might Like This": Exploring Effects of Confidence Signal Patterns on Trust in and Reliance on Conversational Recommender Systems. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 792–804. https://doi.org/10.1145/3593013.3594043

[104] T Venkat Narayana Rao, Akhila Gaddam, Muralidhar Kurni, and K Saritha. 2022. Reliance on artificial intelligence, machine learning and deep learning in the era of industry 4.0. *Smart healthcare system design: security and privacy aspects* (2022), 281–299.

[105] 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. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (March 2022), 1–22. https://doi.org/10.1145/3512930

[106] Lauren Rhue. 2019. Beauty's in the AI of the Beholder: How AI Anchors Subjective and Objective Predictions. *ICIS 2019 Proceedings* (Nov. 2019). https://aisel.aisnet.org/icis2019/future_of_work/future_work/15

[107] Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. 2022. Understanding the Role of Explanation Modality in AI-assisted Decision-making. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. ACM, Barcelona Spain, 223–233. https://doi.org/10.1145/3503252.3531311

[108] Imre J Rudas and János Fodor. 2008. Intelligent systems. *International Journal of Computers, Communications & Control* 3, 3 (2008), 132–138.

[109] William Samuelson and Richard Zeckhauser. 1988. Status quo bias in decision making. *Journal of risk and uncertainty* 1 (1988), 7–59.

[110] 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), 695–720.

[111] Nicolas Scharowski, Sebastian A. C. 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 (July 2023), 1151150. https://doi.org/10.3389/fcomp.2023.1151150

[112] 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. *arXiv preprint arXiv:2310.02108* (2023).

[113] Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl, and Michael Vössing. 2022. A Meta-Analysis of the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making. 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, 617–626. https://doi.org/10.1145/3514094.3534128

[114] 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*. ACM, Sydney NSW Australia, 410–422. https://doi.org/10.1145/3581641.3584066

[115] Anuschka Schmitt, Thiemo Wambsganss, Matthias Soellner, and Andreas Janson. 2021. Towards a Trust Reliance Paradox? Exploring the Gap Between Perceived Trust in and Reliance on Algorithmic Advice. *ICIS 2021 Proceedings* (Dec. 2021). https://aisel.aisnet.org/icis2021/ai_business/ai_business/14

[116] Thomas B Sheridan, Thomas B Sheridan, Kybernetiker Maschinenbauingenieur, Thomas B Sheridan, and Thomas B Sheridan. 2002. *Humans and automation: System design and research issues*. Vol. 280. Human Factors and Ergonomics Society Santa Monica, CA.

[117] Joon Gi Shin, Janin Koch, Andrés Lucero, Peter Dalsgaard, and Wendy E Mackay. 2023. Integrating AI in Human-Human Collaborative Ideation. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–5.

[118] 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.

[119] Elizabeth Solberg, Magnhild Kaarstad, Maren H Rø Eitrheim, Rossella Bisio, Kine Reegård, and Marten Bloch. 2022. A conceptual model of trust, perceived risk, and reliance on AI decision aids. *Group & Organization Management* 47, 2 (2022), 187–222.

[120] Divya K. Srivastava, J. Mason Lilly, and Karen M. Feigh. 2022. Improving Human Situation Awareness in AI-Advised Decision Making. In *2022 IEEE 3rd International Conference on Human-Machine Systems (ICHMS)*. IEEE, Orlando, FL, USA, 1–6. https://doi.org/10.1109/ICHMS56717.2022.9980783

[121] 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 (Dec. 2022), 491–508. https://doi.org/10.1007/s42113-022-00157-y

[122] S. Tolmeijer, M. Christen, S. Kandul, M. Kneer, and A. Bernstein. 2022. Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making. In *Conf Hum Fact Comput Syst Proc*. Association for Computing Machinery. https://doi.org/10.1145/3491102.3517732 Journal Abbreviation: Conf Hum Fact Comput Syst Proc.

[123] Eric Lansdown Trist and Kenneth W Bamforth. 1951. Some social and psychological consequences of the longwall method of coal-getting: An examination of the psychological situation and defences of a work group in relation to the social structure and technological content of the work system. *Human relations* 4, 1 (1951), 3–38.

[124] 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. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 1–38. https://doi.org/10.1145/3579605

[125] Polyxeni Vassilakopoulou, Elena Parmiggiani, Arisa Shollo, and Miria Grisot. 2022. Responsible AI: Concepts, critical perspectives and an Information Systems research agenda. (2022).

[126] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[127] Mor Vered, Tali Livni, Piers Douglas Lionel Howe, Tim Miller, and Liz Sonenberg. 2023. The effects of explanations on automation bias. *Artificial Intelligence* 322 (Sept. 2023), 103952. https://doi.org/10.1016/j.artint.2023.103952

[128] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*. ACM, College Station TX USA, 318–328. https://doi.org/10.1145/3397481.3450650

[129] Christopher D Wickens. 1995. Designing for situation awareness and trust in automation. *IFAC Proceedings Volumes* 28, 23 (1995), 365–370.

[130] 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 (March 2023), 103839. https://doi.org/10.1016/j.artint.2022.103839

[131] Qiaoning Zhang, Matthew L Lee, and Scott Carter. 2022. You Complete Me: Human-AI Teams and Complementary Expertise. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–28.   https://doi.org/10.1145/3491102.3517791

[132] Jieqiong Zhao, Yixuan Wang, Michelle V. Mancenido, Erin K. Chiou, and Ross Maciejewski. 2023. Evaluating the Impact of Uncertainty Visualization on Model Reliance. *IEEE Transactions on Visualization and Computer Graphics* (2023), 1–15.   https://doi.org/10.1109/TVCG.2023.3251950

[133] Ingrid Zukerman, Andisheh Partovi, and Jakob Hohwy. 2023. Influence of Device Performance and Agent Advice on User Trust and Behaviour in a Care-taking Scenario. *User Modeling and User-Adapted Interaction* 33, 5 (Nov. 2023), 1015–1063.   https://doi.org/10.1007/s11257-023-09357-y