

A Meta-Analysis of the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making

Max Schemmer*
max.schemmer@kit.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

Patrick Hemmer*
patrick.hemmer@kit.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

Maximilian Nitsche
maximilian.nitsche@student.kit.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

Niklas Kühn
niklas.kuehl@kit.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

Michael Vössing
michael.voessing@kit.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

ABSTRACT

Research in artificial intelligence (AI)-assisted decision-making is experiencing tremendous growth with a constantly rising number of studies evaluating the effect of AI with and without techniques from the field of explainable AI (XAI) on human decision-making performance. However, as tasks and experimental setups vary due to different objectives, some studies report improved user decision-making performance through XAI, while others report only negligible effects. Therefore, in this article, we present an initial synthesis of existing research on XAI studies using a statistical meta-analysis to derive implications across existing research. We observe a statistically positive impact of XAI on users' performance. Additionally, the first results indicate that human-AI decision-making tends to yield better task performance on text data. However, we find no effect of explanations on users' performance compared to sole AI predictions. Our initial synthesis gives rise to future research investigating the underlying causes and contributes to further developing algorithms that effectively benefit human decision-makers by providing meaningful explanations.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → *Artificial intelligence*.

KEYWORDS

Explainable Artificial Intelligence, Decision-Making, Empirical Studies, Meta-Analysis

ACM Reference Format:

Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühn, and Michael Vössing. 2022. A Meta-Analysis of the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making. In *Proceedings of the 2022*

*Both authors contributed equally to this research.

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

AIES'22, August 1–3, 2022, Oxford, United Kingdom

© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9247-1/22/08...\$15.00

<https://doi.org/10.1145/3514094.3534128>

AAAI/ACM Conference on AI, Ethics, and Society (AIES'22), August 1–3, 2022, Oxford, United Kingdom. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3514094.3534128>

1 INTRODUCTION

Over the last years, the rapid developments in artificial intelligence (AI) have increased its use in many application domains. In this context, AI's continuously rising capabilities have surpassed human performance in an increasing number of tasks, such as playing poker [7], go [56], or correctly recognizing various categories of interest in images [24]. Due to these remarkable developments, AI is increasingly applied to support decision-makers in an increasing number of domains, such as medicine [42, 63], finance [16], law [33] or manufacturing [57].

To offer decision-makers meaningful support, AI models are expected to provide accurate predictions and a notion of how a particular decision has been derived. In particular, explaining the rationale behind an algorithmic decision should enable domain experts to learn when to trust the recommendations of the AI and when to question it [67]. This requirement fueled the continuous development of explainability techniques from the field of explainable AI (XAI), intending to make the decision-making process of black-box AI models more transparent and, thus, comprehensible for domain experts [1]. Common approaches include among others feature importance-based [49], example-based [10], or rule-based methods [50]. A better understanding of how the AI's decision was derived should subsequently enable the user to appropriately rely on the AI's suggestions on a case-by-case basis [4, 38]. For instance, explanations contradicting the AI's prediction could signify the user to become skeptical, consequently considering the AI prediction less in the final decision-making process.

With the ongoing development of XAI techniques, researchers have started to evaluate AI with and without explanations to assess whether their utility for better decision-making can be quantified [4, 8, 9, 12, 14, 22, 23, 27, 36, 37, 39, 60, 65, 67]. Whereas some researchers identify a benefit of XAI-based decision support in user studies [8, 36], others find only negligible evidence [12, 39], with the underlying causes remaining partly unexplored [28, 54]. Therefore, in this article, we aim to clarify the current "snapshot" of the utility of XAI-based decision support. We conduct a meta-analysis of user studies identified in a structured literature review to shed light on the effect of XAI-assisted decision-making on user performance.

In detail, our analysis encompasses studies that allow a comparison between human, AI-, and XAI-assisted task performance. Our initial findings are the following: First, on average, XAI-assisted decision-making enhances human task performance compared to no assistance at all. However, we find no additional effect of explanations on users' performance in XAI-assisted decision-making compared to isolated AI predictions, which raises questions on how to further develop current XAI methods that improve users' task performance. Second, we find that distinct data types affect user performance differently. In this context, human-AI decision-making turns out to be more effective on text data compared to tabular data.

The remainder of this article is structured as follows. In Section 2, we first outline related work in the context of human-AI decision-making. In Section 3, we describe the methodological approach of our meta-study. Subsequently, we present the results of the meta-study, including a subgroup analysis in Section 4. In this context, we provide additional qualitative insights on an individual level. We outline the current limitations of our work in Section 5, followed by a discussion on relevant implications that result from these findings for the future development of XAI algorithms in Section 6. Finally, Section 7 concludes our work.

2 RELATED WORK

Over the last years, research has focused on developing algorithms that provide explanations for AI predictions [1, 15]. By now, these algorithms are increasingly employed in a growing number of practical use cases such as in manufacturing [55, 59], medicine [47], or the hospitality industry [62]. Usually, XAI is utilized in scenarios that involve humans-in-the-loop processes. The underlying idea is that humans will benefit from the AI's suggestion if it is accompanied with an explanation. Therefore, a constantly rising number of studies has started to analyze the effects of explanations in behavioral experiments [28]. In these experiments, many different target variables are taken into consideration, e.g., whether humans are capable of predicting what a model would recommend (proxy tasks) [8, 13, 23] or whether explanations support them in model debugging [2, 32].

In the scope of this study, we explicitly focus on AI-assisted decision-making—a setting in which an AI supports a human with the goal of improving the decision-making quality. The prediction of the AI might be accompanied by additional information, e.g., about its prediction uncertainty or different types of explanations. After receiving the AI's advice, the human decision-maker is responsible for making the final decision. A scenario, which is often also required from a legal perspective, as the human needs to make the final decision [5]. By providing either additional information on the AI's prediction uncertainty [43, 67] or explanations on how a decision was derived [4, 37], humans shall be enabled to better question the AI's decision. To develop a deeper understanding of this assumption, research has evaluated the effect of explanations on users' trust and how reliance on AI decisions can be appropriately calibrated [4, 9, 34, 52, 66, 67]. In this context, providing humans not only with the AI's prediction and respective explanations but also with a notion about its global performance can influence the overall team performance [37]. Additional benefits can be found when humans are provided with model-driven tutorials about AI

functionality and the task itself [36]. Further work in this line of research has investigated the influence of AI advice in the out-of-distribution setting—instances differing from the distribution used for AI training—on the final human decision [39].

Besides these factors, the explanation type of an AI prediction can play a decisive role. In this context, research has developed various explainability techniques [1] ranging from feature importance methods [49] over example-based approaches [10] to rule-based explanations [50] that have been evaluated in user studies accordingly [28]. However, the current picture emerging from the results of different studies regarding the effects of XAI methods on AI-assisted decision-making performance is not unambiguous. Whereas, e.g., Carton et al. [12] conclude that feature-based explanations do not help users in classification tasks, Hase and Bansal [23] find some of them to be effective in model simulatability, which refers to the ability to predict the model behavior given an input and an explanation. In this context, further studies demonstrate the utility of explanations [8], whereas others find that they can convince humans to follow incorrect suggestions more easily [4, 60].

Of course, ambiguous findings can also be attributed to the specific setups of each study and the different goals pursued by the researchers. We aim to shed light on this ambiguity by conducting a meta-analysis of human-AI decision-making—particularly on the influence of explainability.

3 METHODOLOGY

We elaborate on our data collection approach to identify relevant articles, followed by the statistical analysis conducted on the final set of user studies.

3.1 Data Collection

For the collection of empirical user studies in the field of XAI, we conducted a structured literature review based on the methodology outlined by vom Brocke et al. [61]. In detail, we developed a search string focusing on XAI and behavioral experiments. For both topics, several synonyms were included after an explorative search. Subsequently, the search string was iteratively refined, resulting in the following final search string:

```
TITLE-ABS-KEY("explainable artificial intelligence" OR XAI OR
"explainable AI" OR ( ( interpretability OR explanation ) AND (
"artificial intelligence" OR AI OR "machine learning" ) ) ) AND (
"human performance" OR "human accuracy" OR "user study" OR
"empirical study" OR "online experiment" OR "human experiment"
OR "behavioral experiment" OR "human evaluation" OR "user
evaluation")
```

To ensure comprehensive coverage of relevant articles, we chose the SCOPUS database for our initial search [53]. We filtered identified articles according to the following three criteria: an article identified with the search string was included if it (a) conducted at least one empirical user study and (b) reported the task performance as a performance measure for humans and AI- or XAI-assisted decision-making on the same task.

Additionally, we conducted a forward and backward search starting from the articles that fulfill our inclusion criteria. We extracted all individual treatments and outcomes for each article. For instance, if an experiment compared AI- with XAI-assisted decision-making

Source	Dataset	Datatype	Studies
Alufaisan et al. [3]	COMPAS [48]	Tabular	AI-, XAI-assisted (Anchor)
	Census [19]		AI-, XAI-assisted (Anchor)
Bansal et al. [4]	LSAT [58]	Text	XAI-assisted (Confidence, Explain Top-1 Expert, Explain Top-2 Expert, Adaptive Expert)
	Book reviews [25]		XAI-assisted (Confidence, Explain Top-1 AI, Explain Top-2 AI, Adaptive AI, Adaptive Expert)
	Beer reviews [41]		XAI-assisted (Confidence, Explain Top-1 AI, Explain Top-2 AI, Adaptive AI, Adaptive Expert)
Buçinca et al. [8]	Fat content prediction [8]	Image	AI-, XAI-assisted (Inductive & Deductive Explanation)
Carton et al. [12]	Online toxicity [64]	Text	AI-, XAI-assisted (Keyword, Partial, Full Explanation)
Fügener et al. [21]	Dog breed ImageNet [51]	Image	AI-, XAI-assisted (Certainty)
Lai et al. [36]	Deception detection [45, 46]	Text	XAI-assisted (Signed & Predicted Label, Signed & Predicted Label & Guidelines, Signed & Predicted Label & Guidelines & Accuracy)
Liu et al. [39]	COMPAS [48]	Tabular	XAI-assisted (Static/Static, Interactive/Static, Interactive/Interactive)
	ICPSR [44]		XAI-assisted (Static/Static, Interactive/Static, Interactive/Interactive)
	BIOS [17]	Text	XAI-assisted (Static/Static, Interactive/Static, Interactive/Interactive)
van der Waa et al. [60]	Diabetes mellitus type 1 [60]	Tabular	AI-, XAI-assisted (Rule-based, Example-based)
Zhang et al. [67]	Census [19]	Tabular	AI-, XAI-assisted (Confidence, Feature Importance)

Table 1: Overview of articles that were identified in the structured literature review and are analyzed in this work. All articles are peer-reviewed at the time of the meta-study.

in a between-subject design in two separate treatments, each of them was registered as a separate record in our database. If an article includes multiple experiments, we perform the data extraction process for each experiment separately. We contacted authors by email in case of missing or not reported information in the articles regarding the conducted user studies.

The collected studies vary considerably in terms of tasks, problem settings, and reported performance metrics. Accordingly, we filter our set of studies in the following way: First, we focus on studies assessing classification tasks as they account for the largest subset across all entries in our database. Second, we restrict the subset of relevant studies to those reporting the mean accuracy as the performance measurement in each study since we require a common metric across multiple studies. This ensures that we base our meta-analysis on comparable and interpretable effect sizes. Third, we only include studies that have been conducted as a between-subject design. By excluding studies conducted in a within-subject design, we avoid taking into account the learning effect of participants between treatments that might distort the effect sizes of our analysis.

Subsequently, we extract all necessary performance metrics from the articles. Next, we define the case in which the human performs the task without any AI support as human performance. If the human is additionally equipped with AI advice, but without explanations, we call the performance AI-assisted performance. AI assistance with explanations is called XAI assistance, and the resulting performance measure is denoted as XAI-assisted performance. Based on these definitions, we excluded all studies that do not report human performance and either AI-assisted or XAI-assisted performance. Based on the resulting sample, we conduct the following statistical analysis.

3.2 Statistical Analysis

For each study, we calculate the effect size as the between-group standardized mean difference (SMD) of the task performance. Furthermore, we report Hedges' g [26] to correct the SMD for a possible upward bias of the effect size when the sample size of a study is small ($n \leq 20$). Thus, Hedges' g is smaller for $n \leq 20$ than the uncorrected SMD but approximately the same for larger sample sizes. We obtain the standard deviations from standard errors and confidence intervals for group means reported for each treatment following

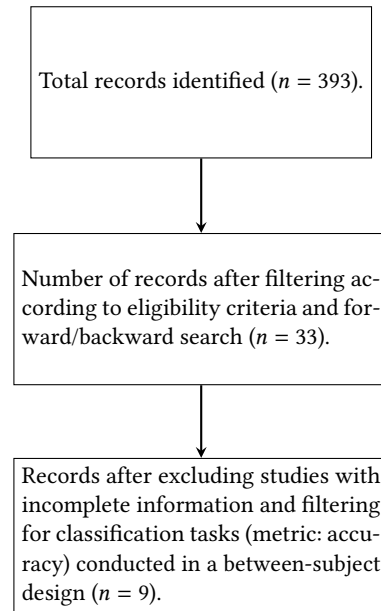


Figure 1: Flowchart describing the data collection and article selection procedure.

the procedure outlined in Higgins et al. [29]. In case an article encompasses multiple studies with a single control group, we divide the size of this control group by the number of studies to avoid multiple comparisons against the same group [29].

For our meta-analytic model and the pooling of effect sizes, we estimate a random-effects model as the setups and populations are considerably heterogeneous between studies. Hence, we calculate the distribution mean of effect sizes instead of estimating and assuming one single true effect size underlying the studies (fixed-effect model [6]). To assess the between-study heterogeneity variance τ^2 and its confidence intervals we use the DerSimonian-Laird estimator [18] and Jackson's method [31], respectively.

Additionally, we conduct a subgroup analysis to provide further insights across current XAI studies. Several studies have discussed that task choice has a strong influence on the experimental outcome [21, 35]. This article focuses on the influence of the task's data type.

In this context, many researchers have argued about the importance of data types in human-AI decision-making. For example, Fügener et al. [21] reason that image recognition, in general, is well suited for human-AI decision-making since it is an intuitive task for humans.

4 RESULTS

We start by presenting the final set of included studies, then outline the meta-study results, including the respective subgroup analyses. Finally, we provide additional qualitative insights on an article level.

4.1 Data Collection

As of February 2022, we identified a total number of 393 articles. After applying our inclusion criteria and conducting a forward and backward search, the number of relevant articles is reduced to 33.

As classification tasks form the largest subset, we focus on this particular prediction problem. After filtering for accuracy as a common metric and removing articles with missing information, e.g., sample size or dispersion measures, we include 9 articles in the meta-analysis and the respective subgroup analyses. Figure 1 visualizes the entire filtering process. Moreover, Table 1 provides an overview of all included articles together with information about each dataset, datatype, and treatments extracted from the articles. Each article contains at least one behavioral experiment conducted with a particular dataset. Each experiment consists of several experimental treatments. The treatment in which humans conducted a task on their own without AI assistance is referred to as a control group. In the following, we denote each treatment as an individual study. Overall, we thereby have a sample size of 44 studies.

4.2 AI Assistance vs. XAI Assistance

We start our meta-analysis by investigating AI- and XAI-assisted performance. For this reason, we first focus on all studies that report AI- and XAI-assisted performance, which leads us to a sample of 11 studies and a total number of 999 observations.

Figure 2 displays the forest plot of the standardized mean difference (SMD) between AI-assisted and XAI-assisted performance. The results of the analysis reveal that, on average, the SMD of all studies that reported AI- and XAI-assisted performance is 0.07 with a 95% confidence interval (CI) [-0.15, 0.30]. A z-test against the null-hypothesis that the effect size is 0 cannot be rejected ($z = 0.63, p = 0.53$). This means we do not find a significant difference between AI-assisted and XAI-assisted performance in our current sample of studies. Regarding heterogeneity, we find an I^2 of 57.00% (95% CI [15.70%, 78.00%]), which can be considered moderate [29]. The τ^2 is 0.07 (95% CI [0.01, 0.54]) and Q is significantly different from 0 ($Q = 23.24, df = 10, p < 0.01$). Thus, we can reject the null hypothesis that the true effect size is identical in all studies. To provide an intuitive understanding of the heterogeneity, we also report the prediction interval that represents the expected range of true effects in other studies [30]. The prediction interval ranges from -0.59 to 0.74. That means we can expect negative as well as positive effects of XAI assistance in comparison to AI assistance. In summary, on average, XAI-assisted decision-making does not significantly influence the performance of human-AI decision-making in our sample. The highest improvement was measured by van der

Waa et al. [60]. Contrary, the highest negative impact of XAI is measured by Alufaisan et al. [3].

4.3 Human vs. XAI Assistance

We analyze the overall effect of XAI in comparison with human performance. Therefore, we filter all studies that report human and XAI-assisted performance. This results in a sample of 33 studies and a total number of 5,083 participants. Based on this sample, we analyze whether XAI-assisted decision-making improves performance compared to humans conducting a task alone.

Figure 3 visualizes the forest plot of the standardized mean difference between human and XAI-assisted performance. The meta-analysis indicates that, on average, XAI assistance increases task performance by 0.59 SMD as compared to humans conducting the tasks alone. The 95% CI of the SMD ranges from 0.39 to 0.79. As this range does not include an effect size of 0 and a z-test is significant ($z = 5.73, p < 0.0001$), we can reject the null hypothesis concluding that, on average, XAI-assisted decision-making improves human task performance. Looking at heterogeneity, we can reject the null hypothesis that the true effect size is identical in all studies ($Q = 219.24, df = 32, p < 0.0001$). Moreover, I^2 is 85.40% with a 95% CI of 80.50% to 89.10%. The estimated τ^2 is 0.29 (95% CI [0.18, 0.59]). Thus, the level of heterogeneity can be considered substantial [29]. The prediction interval in the analysis is -0.54 to 1.71. This means that we cannot say with certainty that XAI always has a positive impact on human decision-making as the prediction interval is not exclusively larger than 0. Even though most studies report a performance improvement through XAI assistance, we also find studies that report a performance decline. In this context, one of the two studies conducted by Alufaisan et al. [3] encountered the most negative effects with a human performance decline. Participants are asked to predict whether a defendant will recidivate in two years and receive AI predictions with decision rules aiming to support users' understanding of the AI's decisions. However, this reduction is not statistically significant due to a high level of dispersion. The most considerable improvement can be found in a study by Buçinca et al. [8]. Here, participants have to decide based on an image of a meal whether the fat content of this meal on a food plate exceeds a certain threshold.

It is important to highlight that the significant SMD does not imply that including explanations will improve performance over simply providing AI advice without any form of explainability, as we did not find a significant difference between AI-assisted and XAI-assisted performance in Section 4.2. Instead, it can be interpreted as a positive effect of *some* form of AI advice.

4.4 Tabular vs. Text Data

Additionally, we conduct a subgroup analysis based on three data types used in our sample—tabular, text, and image data. As only two articles report experiments using image data, the interpretation of this data type is not conclusive. Thus, we focus on tabular and text data types for the subgroup analysis resulting in a total sample size of 31 studies and 4,702 participants. Figure 4 displays the forest plot of the standardized mean difference between human and XAI-assisted performance with regard to both data types.

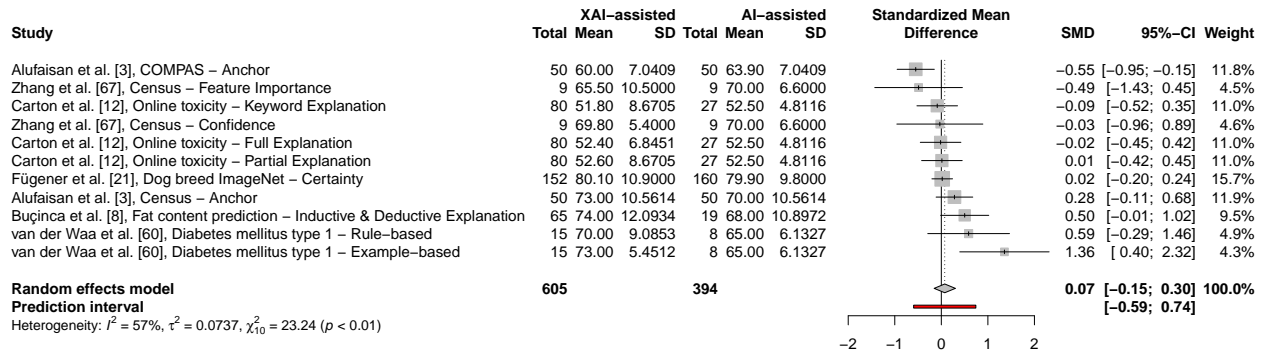


Figure 2: Forest plot of the standardized mean difference between AI-assisted and XAI-assisted performance.

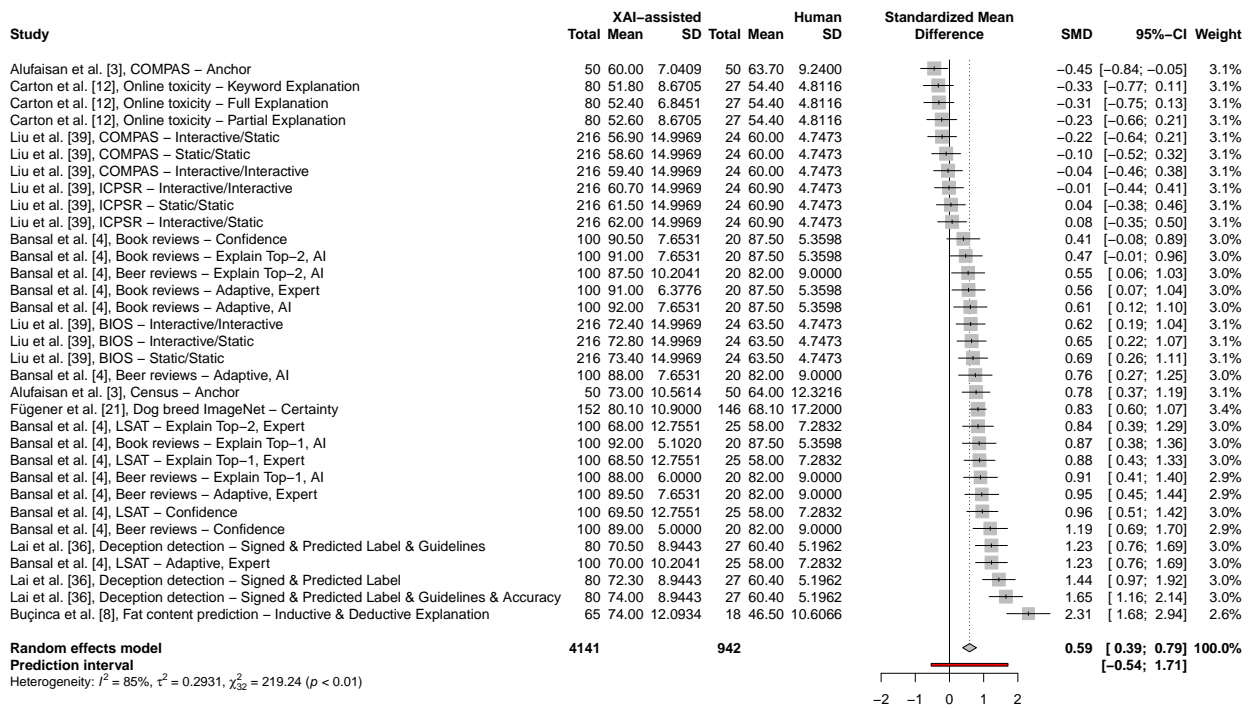


Figure 3: Forest plot of the standardized mean difference between human and XAI-assisted performance.

The SMD of the tabular subgroup is 0.01 (95% CI [-0.24, 0.26]). As this range does include an effect size of 0 and the z-value is 0.08 with a corresponding p-value of 0.94, we cannot reject the null hypothesis. That means we cannot conclude that the SMD is significantly different from 0. Regarding heterogeneity, we can reject the null hypothesis that the true effect size is identical in all studies ($Q = 20.30$, $df = 7$, $p < 0.005$). The I^2 is 65.50% with a 95% CI ranging from 26.70% to 83.80% and τ^2 has a value of 0.09 (95% CI [0.01, 0.48]). Finally, the prediction interval ranges from -0.77 to 0.79. That means we can expect future negative impacts of XAI assistance on human decision performance in certain situations.

The text data subgroup has a higher SMD of 0.72 (95% CI [0.50, 0.93]). This range does not include an effect size of 0. Additionally, the z-value is 6.65 with $p < 0.0001$ denoting that the SMD is significantly different from 0. In terms of heterogeneity, we can reject the null hypothesis that the true effect size is identical in all studies ($Q = 103.57$, $df = 22$, $p < 0.0001$). In this context, τ^2 is 0.21 (95% CI [0.10, 0.47]) and I^2 is 78.80% (95% CI [68.70%; 85.60%]). Thus, the heterogeneity of the text data subgroup can be considered higher than the tabular data subgroup. The prediction interval ranges from -0.26 to 1.69, which means we can also expect some negative XAI effects with text data.

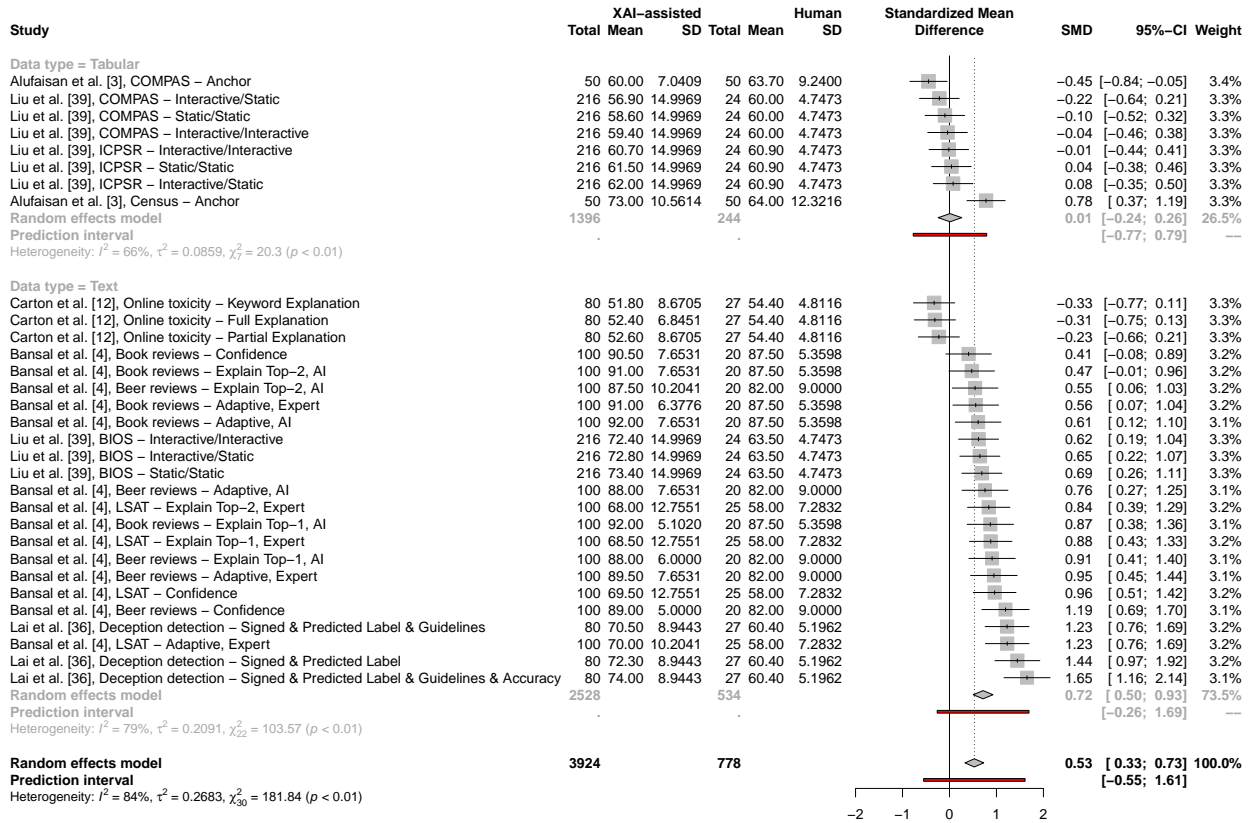


Figure 4: Forest plot of the standardized mean difference between human and XAI-assisted performance considering the subgroups tabular and text data.

Lastly, comparing both subgroups, we observe significant performance differences between tabular and text data suggesting that the data type in our sample influences the effect of XAI assistance on the performance ($Q = 17.81, df = 1, p < 0.0001$).

4.5 Summary of the Articles

Having analyzed the collected studies in the form of a meta-analysis, we pursue a discussion and summarization of the individual articles from a qualitative perspective. We focus on extracting further insights that can be derived from comparing human performance without any assistance and AI or XAI assistance. In particular, we are interested in the question of why XAI assistance did or did not improve AI assistance.

Alufaisan et al. [3] draw upon rule-based explanations generated by anchor LIME [50] for two real-world tabular datasets—an income prediction task using the Census dataset [19] and a recidivism prediction task using the COMPAS dataset [48]. The XAI algorithm provides rules which are denoted as anchors to explain a prediction on an instance level. A rule is considered an anchor if any changes in the features that are not included in the anchor do not impact the AI decision [3]. The authors find that AI assistance improves human performance. However, they find no additional benefit of

XAI assistance regarding decision-making performance. They hypothesize that one reason for the negligible effects of XAI could be information overload. When taking a closer look at the individual decision level, humans tend to follow AI predictions more often when the AI makes a correct prediction than when it makes an incorrect one. In this context, explanations did not alter this observation.

Bansal et al. [4] compare multiple forms of explanations of AI predictions on three tasks. Two of them are about sentiment analysis of book [25], and beer reviews [41]. The third task consists of a set of law school admission test questions requiring logical reasoning [58]. All three datasets consist of text data. As explanations, they use feature importance generated by LIME [49] and feature importance generated by human experts. Additionally, the authors evaluate the effect of providing only confidence ratings of the AI’s prediction, which we consider as a form of explanation as well to ensure that AI assistance encompasses in all studies only the predictions of the AI. Therefore, we are restricted to the comparison of human performance with XAI assistance in the meta-analysis. Interestingly, the authors find no performance improvements of explanations over confidence ratings. However, they find that providing explanations tends to increase the chance that humans accept the AI’s prediction regardless of its correctness—which shows contrary results to the

observations of Alufaisan et al. [3]. Moreover, participants receiving XAI assistance outperform both participants and AI alone.

Buçinca et al. [8] analyze two different techniques for evaluating XAI systems. First, a proxy task, e.g., asking humans to predict the AI's decision. Second, an actual decision-making task, e.g., asking humans to make a decision with XAI assistance. We consider the actual decision-making task for the meta-analysis. The authors show participants images of different food plates and ask them to predict the fat content of the meal. The participants are requested to decide whether the fat content in percent is higher than a threshold. Users are provided with inductive (i.e., example-based) and deductive (i.e., rule-based) explanations. With XAI assistance, the users significantly outperform users with AI assistance, which yields better performance than users conducting the task without any support. Interestingly, example-based explanations enable users to identify erroneous AI predictions better. However, they prefer and trust the rule-based explanations more.

Carton et al. [12] consider a social media comment toxicity prediction task for their experiment by sampling comments from the dataset provided by Wulczyn et al. [64]. They utilize feature importance by highlighting passages and words pointing towards a toxic comment. In detail, three different forms of explanations are analyzed—full, partial, and keyword explanations. The full explanations are generated by an attention model [11] that produces a discrete attention mask over the input text for toxic content. Partial explanations refer to reducing the mask to the most toxic passage. Keyword explanations are derived from a bag-of-words logistic regression classifier that highlights only the most toxic single words instead of considering the context. Their study shows a marginal negative trend of AI assistance performance in comparison to users with no assistance at all. However, the effect does not vary significantly regardless of explanations being present or not. In this context, users tend to follow the AI prediction without an effect of the different explanations provided. A detailed analysis of false negative and false positive rates reveals that explanations tend to increase false negatives and reduce false positives. The authors hypothesize that this might indicate a reduced cognitive engagement with the social media comments by focusing on the highlighted text without considering the not highlighted passages.

Fügener et al. [21] use a dog breed image classification task. They evaluate the effect of AI assistance with and without additional confidence ratings and compare both treatments with the performance of users that do not receive any AI support. While the authors find that AI assistance significantly improves human performance, no additional effect of providing confidence ratings can be identified. In this context, both AI-assisted performance and the one with additional confidence ratings outperform humans and AI when conducting the task alone. Interestingly, a detailed analysis of the confidence rating treatment reveals that providing AI certainty decreases AI adherence. The authors hypothesize that a decrease in users' trust might explain this effect as it declines in this treatment as well.

Lai et al. [36] consider a deception detection task which is about identifying fake hotel reviews [45, 46]. The authors analyze the effect of three types of XAI assistance provided together with the AI prediction. First, they provide users with signed feature importance, i.e., they highlight the most important words that indicate real and

fake reviews. Word importance is derived from the absolute value of the coefficients using a linear SVM with a unigram bag-of-words. Second, they additionally provide the participants with supporting guidelines derived from related research and observations of the model, which are paraphrased by the authors. Third, they provide an additional AI accuracy performance statement on top. Before the actual decision-making task, participants had to undergo a task-specific training phase. In a prior experiment, the authors demonstrated that these so-called tutorials enhance participants' performance without any further assistance from the AI. In general, they find an improvement in XAI-assisted performance over human performance in all three treatments. However, no statistical significance between the different forms of XAI assistance can be detected. Moreover, the authors also emphasize that XAI-assisted performance remains inferior to the AI performing the task alone, even though a small proportion of participants can be identified that were able to outperform the AI. The authors do not consider an AI-assisted treatment in their experiment.

Liu et al. [39] explore the effect of out-of-distribution data instances and interactive explanations on human-AI decision-making. For the meta-analysis, we focus on the in-distribution setting, including a comparison of static with interactive explanations, as out-of-distribution data might impede the comparability with the other studies. The authors draw upon three tasks in their article. In the first, participants predict whether arrested defendants will violate the terms of pretrial release using the ICPSR dataset [44]. The second task is about predicting whether defendants will recidivate in two years using the COMPAS dataset [48]. Both are tabular datasets. The third task requires participants to predict a person's profession given a textual biography using the BIOS dataset [17]. Regarding the AI model, a linear SVM classifier with unigram bag-of-words for BIOS and one-hot encoded features for ICPSR and COMPAS is employed. The static explanations consist of feature importance by coloring features that contribute to the AI prediction. Interactive explanations offer users the possibility to explore what-if scenarios. In general, the authors find that XAI assistance improves users' performance in predicting a person's profession compared to users without any assistance. However, no significant performance difference between static and interactive explanations can be found. For the two recidivism prediction datasets, they find no significant difference between the performance of XAI assistance and human alone. The authors hypothesize that the complexity of both recidivism tasks might have prevented noticeable performance improvements. Interestingly, even though no performance difference can be observed, users rate interactive explanations more useful in the recidivism prediction tasks. The authors do not consider sole AI-assisted decision-making in their article.

van der Waa et al. [60] evaluate the effect of example- and rule-based explanations in the context of a diabetes self-management task where participants are requested to select the appropriate dose of insulin. The authors compare AI assistance with both types of explanations but do not report sole human performance. In this context, the presence of either example- or rule-based explanations does not result in a significant performance difference compared to pure AI assistance. A closer analysis of explanations' "persuasive power", i.e., how often humans agree with the AI recommendation regardless of being correct or not, reveals that users without

explanations follow the AI prediction significantly less than with example- or rule-based explanations.

Zhang *et al.* [67] compare the effect of additional confidence ratings with feature importance explanations using Shapley values [40] in the context of human-AI decision-making. They utilize the Census dataset [19] for asking participants to predict whether a person's income would exceed \$50,000. In the experiment, the authors find no significant difference between additional confidence ratings and feature importance explanations in terms of task performance. Moreover, both treatments do not differ significantly from sole AI assistance. Even though displaying confidence scores does not affect task performance, it can be found that it improves overall trust and contributes to a calibration over different confidence levels. The authors explain this phenomenon with a high correlation between human and AI confidence, showing a large overlap of instances with low AI and human confidence. In contrast, explanations do not affect users' trust in their experiment.

In summary, most studies, except for Bućinca *et al.* [8], found no significant differences when comparing XAI and AI assistance on an individual study level. Alufaisan *et al.* [3] argue that this negligible effect of XAI might stem from information overload. Carton *et al.* [12] discuss that XAI, in the case of feature importance, might reduce the amount of data that humans process as they just focus on the features that are relevant for the AI. Liu *et al.* [39] discuss whether task complexity could be a reason for no significant improvements of XAI assistance over AI assistance. We also observe some contrasting results that require further investigation in future work. Fügener *et al.* [21] find a general decrease in AI adherence in the context of confidence ratings. In contrast, van der Waa *et al.* [60] and Bansal *et al.* [4] find that explanations just increase the general probability of accepting AI advice.

Based on the qualitative review, we derive two potential factors influencing the utility of XAI: First, XAI could improve decision-making performance by increasing the acceptance of AI advice. Note that the performance improvement will just happen if the AI, on average, performs significantly better than the human [67]. Second, XAI assistance could influence appropriate trust and reliance, which means humans can discriminate between correct and incorrect AI advice [4]. The idea is that humans will be better able to distinguish between correct and incorrect advice if the AI conveys its reasoning. Our qualitative review showed that while there is evidence for increased acceptance of AI advice due to XAI [4, 60, 67], just one article reports some form of appropriate reliance [8].

5 LIMITATIONS

As XAI is a relatively new field of research, at least in comparison to research domains where meta-analyses are more common, e.g., medical research [20], we encounter some major limitations that form around the current existing sample of XAI studies.

First, the current existing sample of XAI studies contains just online studies. In these online studies, people are recruited via online platforms such as Mechanical Turk. They conduct a task not in a controlled lab environment inducing higher variability. Second, the studies use different XAI algorithms ranging from providing an additional confidence score to personalized explanations. We also considered a subgroup analysis for the XAI algorithm category

but are limited due to the current sample size. Therefore, interpretable findings cannot be derived yet. Third, also task design differs between the studies. Some studies use more intuitive tasks for humans, such as sentiment analysis of reviews, while others consider more complicated ones, such as income prediction. Future studies should evaluate other task-related factors beyond data type. Fourth, in the data type subgroup, the tabular subgroup contains just two articles [3, 39]. Even though it contains 8 studies, this poses a possible limitation.

Moreover, as many studies did not report dispersion metrics numerically, we needed to extract them from the plots. However, we conducted a multi-step approach. Two researchers extracted the values individually and afterward discussed the differences. Furthermore, we want to highlight that the meta-analysis is limited with regard to drawing causal conclusions. Our analysis should instead be considered as a synthesis of existing research. The main limitation of the comparison of AI-assisted and XAI-assisted performance is the small sample size due to many studies not reporting the dispersion of their AI-assisted condition.

6 DISCUSSION AND FUTURE WORK

In this work, we conducted the first meta-analysis of XAI-assisted decision-making. Based on a structured data collection process, we collected 393 XAI-related articles. After applying our inclusion criteria, we identified a set of 9 articles encompassing 44 studies for the meta-analysis.

In the current sample, we find no statistically significant difference between XAI- and AI-assisted performance. In this context, we observed that some studies reported positive XAI assistance performance effects, whereas others found no effects or even slightly negative effects.

Additionally, we find a positive effect of XAI assistance on human task performance. Since we do not identify a difference between XAI and AI assistance, the results need to be interpreted carefully. Therefore, we cannot conclude that XAI will lead to an overall performance superior to AI assistance. However, we observe a positive tendency of AI to support humans in decision-making, either with explanations or without. A promising avenue for future research is to investigate the factors that determine consistent performance gains in human-AI decision-making. In this context, the work of Lai *et al.* [35] could be a foundation for the data collection.

Furthermore, our subgroup analysis indicates a stronger positive effect on task performance with text data than tabular data. If this effect can be confirmed in future studies, more work would be required regarding human-AI decision-making with tabular data. Reasons for this difference could be that text data is a more intuitive data type for humans. Additionally, analyzing the utility of explanations and AI predictions concerning image data requires more attention, as we found a total of only 2 articles that used image data. Future work should explicitly investigate performance differences induced by different data types.

Moreover, due to the currently high heterogeneity of the studies, future analyses could consider not only a distinction by data type but also by task type, e.g., complexity, and users' task-specific knowledge. For example, in easy tasks, humans might be able to evaluate better whether AI advice is correct or not.

Our qualitative review of the collected studies highlights that explanations can easily lead to increased acceptance of AI advice. In the scenario in which AI performs, on average, better than humans, its performance might bound the maximum joint performance of both. However, if the goal is that human-AI decision-making ideally results in superior team performance [4], appropriate reliance on the AI's decision becomes indispensable. Therefore, future research should investigate design mechanisms that enable appropriate reliance. Additionally, from this observation emerges the need to discuss the ethical implications of pure acceptance increase in future work as it can be understood as a form of manipulating people to blindly accept AI advice.

7 CONCLUSION

This article presents the results of a meta-analysis of the utility of XAI-assisted decision-making. We identify a total sample of 9 articles that report all necessary information as a prerequisite through a structured literature review. We analyze whether humans' decision-making can benefit from AI support with and without explanations and derive three major findings: First, we do not find a significant effect of state-of-the-art explainability techniques on AI-assisted performance. Second, we observe a significant positive effect of XAI assistance on human performance. Third, our analysis indicates that XAI assistance is more effective on text than tabular data. We hope that our work will motivate scholars to pursue meta-analyses in future human-AI research to systematically assess previous studies to derive conclusions about the current body of research.

REFERENCES

- [1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.
- [2] Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. 2020. Debugging tests for model explanations. *arXiv preprint arXiv:2011.05429* (2020).
- [3] Yasmeen Alufaisan, Laura R Marusich, Jonathan Z Bakdash, Yan Zhou, and Murat Kantarcioglu. 2021. Does explainable artificial intelligence improve human decision-making?. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 6618–6626.
- [4] 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.
- [5] Kevin Bauer, Oliver Hinz, Wil van der Aalst, and Christof Weinhardt. 2021. Expl(AI)n it to me—explainable AI and information systems research. *Business & Information Systems Engineering* 63, 2 (2021), 79–82.
- [6] Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. 2010. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research synthesis methods* 1, 2 (2010), 97–111.
- [7] Noam Brown and Tuomas Sandholm. 2019. Superhuman AI for multiplayer poker. *Science* 365 (2019), 885–890.
- [8] Zana Bućinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 454–464.
- [9] Zana Bućinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.
- [10] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 258–262.
- [11] Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2018. Extractive adversarial networks: high-recall explanations for identifying personal attacks in social media posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.
- [12] Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2020. Feature-based explanations don't help people detect misclassifications of online toxicity. In *Proceedings of the International AAAI Conference on Web and Social Media*. 95–106.
- [13] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. 2018. Do explanations make VQA models more predictable to a human?. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 1036–1042.
- [14] Eric Chu, Deb Roy, and Jacob Andreas. 2020. Are visual explanations useful? A case study in model-in-the-loop prediction. *arXiv preprint arXiv:2007.12248* (2020).
- [15] Arun Das and Paul Rad. 2020. Opportunities and challenges in explainable artificial intelligence (XAI): a survey. *arXiv preprint arXiv:2006.11371* (2020).
- [16] Min-Yuh Day, Tun-Kung Cheng, and Jheng-Gang Li. 2018. AI robo-advisor with big data analytics for financial services. In *International Conference on Advances in Social Networks Analysis and Mining*. 1027–1031.
- [17] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnamurthy Kenchadapadi, and Adam Tauman Kalai. 2019. Bias in bios: a case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 120–128.
- [18] Rebecca DerSimonian and Nan Laird. 1986. Meta-analysis in clinical trials. *Controlled Clinical Trials* 7, 3 (1986), 177–188.
- [19] Dheeru Dua and Casey Graff. 2017. UCI machine learning repository. <http://archive.ics.uci.edu/ml>
- [20] JA Eaden, KR Abrams, and JF Mayberry. 2001. The risk of colorectal cancer in ulcerative colitis: a meta-analysis. *Gut* 48, 4 (2001), 526–535.
- [21] Andreas Fügener, Jörn Grahl, Alok Gupta, and Wolfgang Ketter. 2021. Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *Management Information Systems Quarterly* 45, 3 (2021), 1527–1556.
- [22] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.
- [23] Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: which algorithmic explanations help users predict model behavior?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5540–5552.
- [24] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. *International Conference on Computer Vision* (2015), 1026–1034.
- [25] Ruining He and Julian McAuley. 2016. Ups and downs: modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*. 507–517.
- [26] Larry V Hedges and Ingram Olkin. 1985. *Statistical methods for meta-analysis*. Academic Press.
- [27] Patrick Hemmer, Max Schemmer, Niklas Kühl, Michael Vössing, and Gerhard Satzger. 2022. On the effect of information asymmetry in human-AI teams. *Human-Centered Explainable AI Workshop at the CHI Conference on Human Factors in Computing Systems* (2022).
- [28] Patrick Hemmer, Maximilian Schemmer, Michael Vössing, and Niklas Kühl. 2021. Human-AI complementarity in hybrid intelligence systems: a structured literature review. In *Proceedings of the 25th Pacific Asia Conference on Information Systems*.
- [29] Julian PT Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J Page, and Vivian A Welch. 2019. *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons.
- [30] Joanna IntHout, John PA Ioannidis, Maroeska M Rovers, and Jelle J Goeman. 2016. Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open* 6, 7 (2016).
- [31] Dan Jackson and Jack Bowden. 2016. Confidence intervals for the between-study variance in random-effects meta-analysis using generalised heterogeneity statistics: should we use unequal tails? *BMC Medical Research Methodology* 16, 1 (2016), 1–15.
- [32] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [33] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The Quarterly Journal of Economics* 133, 1 (2018), 237–293.
- [34] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let me explain: impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [35] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-AI decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471* (2021).
- [36] Vivian Lai, Han Liu, and Chenhao Tan. 2020. "Why is 'Chicago' deceptive?" Towards building model-driven tutorials for humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

- [37] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: a case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 29–38.
- [38] John D. Lee and Katrina A. See. 2004. Trust in automation: designing for appropriate reliance. *Human Factors* 46 1 (2004), 50–80.
- [39] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making. *Proceedings of the ACM on Human-Computer Interaction* 5 (2021), 1–45.
- [40] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*.
- [41] Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *12th International Conference on Data Mining*. 1020–1025.
- [42] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg Corrado, Ara Darzi, Mozziyar Etemadi, Florencia Garcia-Vicente, Fiona J. Gilbert, Mark D. Halling-Brown, Demis Hassabis, Sunny Jansen, Alan Karthikesalingam, Christopher J. Kelly, Dominic King, Joseph R. Ledlam, David S. Melnick, Hormuz Mostofi, Lily H. Peng, Joshua Jay Reicher, Bernardino Romera-Paredes, Richard Sidebottom, Mustafa Suleyman, Daniel Tse, Kenneth C. Young, Jeffrey De Fauw, and Shrivya Shetty. 2020. International evaluation of an AI system for breast cancer screening. *Nature* 577 (2020), 89–94.
- [43] Giang Nguyen, Daeyoung Kim, and Anh Nguyen. 2021. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. In *Advances in Neural Information Processing Systems*.
- [44] United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics. 2014. State court processing statistics, 1990–2009: felony defendants in large urban counties.
- [45] Myle Ott, Claire Cardie, and Jeffrey T Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 497–501.
- [46] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies—Volume 1*. 309–319.
- [47] Matteo Pennisi, Isaak Kavasidis, Concetto Spampinato, Vincenzo Schinà, Simone Palazzo, Francesco Rundo, Massimo Cristofaro, Paolo Campioni, Elisa Pianura, Federica Di Stefano, Ada Petrone, Fabrizio Albarello, Giuseppe Ippolito, Salvatore Cuzzocrea, and Sabrina Conoci. 2021. An explainable AI system for automated COVID-19 assessment and lesion categorization from CT-scans. *Artificial Intelligence in Medicine* 118 (2021), 102114–102114.
- [48] ProPublica. 2016. Machine bias: there's software used across the country to predict future criminals. And it's biased against blacks. (2016). <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- [49] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1135–1144.
- [50] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: high-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- [51] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.
- [52] Max Schemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. On the influence of explainable AI on automation bias. In *Proceedings of the 30th European Conference on Information Systems*.
- [53] Michiel Schotten, Wim JN Meester, Susanne Steinginga, Cameron A Ross, et al. 2017. A brief history of Scopus: The world's largest abstract and citation database of scientific literature. In *Research analytics*. Auerbach Publications, 31–58.
- [54] Jakob Schöffer, Maria De-Arteaga, and Niklas Kuehl. 2022. On the relationship between explanations, fairness perceptions, and decisions. *Human-Centered Explainable AI Workshop at the CHI Conference on Human Factors in Computing Systems* (2022).
- [55] Julian Senoner, Torbjørn H. Netland, and Stefan Feuerriegel. 2021. Using explainable artificial intelligence to improve process quality: evidence from semiconductor manufacturing. *Management Science* (2021).
- [56] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, L. Sifre, Dhharshan Kumaran, Thore Graepel, Timothy P. Lillicrap, Karen Simonyan, and Demis Hassabis. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362 (2018), 1140–1144.
- [57] Maximilian Stauder and Niklas Kühl. 2021. AI for in-line vehicle sequence controlling: development and evaluation of an adaptive machine learning artifact to predict sequence deviations in a mixed-model production line. *Flexible Services and Manufacturing Journal* (2021), 1–39.
- [58] LSAT Prep Books Team. 2017. *LSAT prep book study guide: quick study & practice test questions for the Law School Admissions Council's (LSAC) Law school admission test*. Mometrix Test Preparation, Beaumont, TX.
- [59] Alexander Treiss, Jannis Walk, and Niklas Kühl. 2020. An uncertainty-based human-in-the-loop system for industrial tool wear analysis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 85–100.
- [60] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating XAI: a comparison of rule-based and example-based explanations. *Artificial Intelligence* 291 (2021), 103404.
- [61] Jan vom Brocke, Alexander Simons, Björn Niehaves, Kai Riemer, Ralf Plattfaut, and Anne Clevén. 2009. Reconstructing the giant: on the importance of rigour in documenting the literature search process. In *Proceedings of the 17th European Conference on Information Systems*.
- [62] Michael Vössing, Niklas Kühl, Matteo Lind, and Gerhard Satzger. 2022. Designing transparency for effective human-AI collaboration. *Information Systems Frontiers* (2022).
- [63] N. Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanislaw Jastrzebski, Thibault Févry, Joe Katsnelson, Eric Kim, S. Wolfson, Ujas Parikh, Sushma Gaddam, L. Lin, Kara Ho, Joshua D. Weinstein, B. Reig, Yiming Gao, H. Toth, Kristine Pysarenko, A. Lewin, Jiyou Lee, Krystal Airola, E. Mema, Stephanie Chung, Esther Hwang, N. Samreen, S. Kim, L. Heacock, L. Moy, Kyunghyun Cho, and K. Geras. 2020. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Transactions on Medical Imaging* 39, 4 (2020), 1184–1194.
- [64] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*. 1391–1399.
- [65] Arnold Yeung, Shalmali Joshi, Joseph Jay Williams, and Frank Rudzicz. 2020. Sequential explanations with mental model-based policies. *Workshop on Human Interpretability in Machine Learning at International Conference on Machine Learning* (2020).
- [66] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2019. Do I trust my machine teammate? An investigation from perception to decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*.
- [67] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 295–305.