# A Meta-Analysis on 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 Kühl
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, first results might indicate that human-AI decision-making yields 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 to investigate the underlying causes as well as contribute to further development of algorithms that effectively benefit human decision-makers in the form of explanations.

## CCS CONCEPTS

• **Human-centered computing** → Empirical studies in HCI.

## KEYWORDS

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

**ACM Reference Format:**
Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl, and Michael Vössing. 2022. A Meta-Analysis on the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making. In *Proceedings of AAAI/ACM Conference on AI, Ethics, and Society (AIES'22).* 9 pages.

## 1 INTRODUCTION

Over the last years, 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 [53], or correctly recognizing various categories of interest in images [23]. Due to these remarkable developments, AI is increasingly applied to support decision-makers in high-stake scenarios, such as medicine [40, 59], finance [15], law [31] or manufacturing [54]. However, in many high-stake scenarios, full automation is not desirable due to ethical concerns, legal concerns, and the high cost of errors.

To offer decision-makers meaningful support, AI models are not only expected to provide accurate predictions but also a notion of how a particular decision has been derived to foster its understandability. 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 [63]. 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 [47], example-based [10], or rule-based methods [57]. 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, 37]. For instance, explanations contradicting the AI's prediction could be a sign for the user to be skeptical, whereupon the AI prediction might be less considered in the final decision-making process.

Concurrently, with the ongoing development of XAI techniques, researchers have started to evaluate AI with and without explanations in user studies to assess whether their hope for better decision-making can be quantified [8, 9, 11, 13, 21, 22, 35, 36, 38, 57, 63]. Whereas some researchers identify a benefit of XAI-based decision support [4, 8], others find only negligible evidence [42], with the underlying causes remaining partly unexplored [50].

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 relevant 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

*Both authors contributed equally to this research.

*AIES'22, August 1-3, 2022, Oxford, UK*

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 to additionally benefit users' task performance. Second, we find that different 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. Subsequently, in Section 3, we describe the methodological approach of our meta-study. Subsequently, we present the results of the meta-study in general and several subgroup analyses in Section 4. 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 (see Section 5). 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, 14]. By now, these algorithms are increasingly employed in a growing number of practical use cases such as in manufacturing [52, 56] or medicine [45]. Usually, XAI is utilized in scenarios that involve humans-in-the-loop processes. The underlying idea is that humans will additionally benefit from the AI's suggestion if it is provided with an explanation of the prediction. Therefore, a constantly rising number of studies has started to analyze the effects of explanations in behavioral experiments [26]. In these experiments, many different target variables have been taken into consideration, e.g., whether humans are capable of better predicting what a model would recommend (proxy tasks) [8, 12] or whether explanations support them in model debugging [2, 30].

In the scope of this study, we explicitly focus on AI-assisted decision-making—a setting in which a human decision-maker is supported by an AI with the goal to improve the decision 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 recommendation, the human decision-maker is responsible for making the final decision. A scenario, which is oftentimes 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 [42, 63] or explanations on how a decision was derived [4, 36], 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, 32, 49, 61–63]. 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 [36]. Additional benefits can be found when humans are provided with model-driven tutorials with regard to AI functionality, and the task itself [35]. 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 [38].

In addition to factors regarding the setting between humans and AI, the explanation type of an AI prediction can also play a decisive role. In this context, research has developed various explainability techniques [1] ranging from feature importance methods [47] over example-based approaches [10] to rule-based explanations [57] that have been evaluated in user studies accordingly [26].

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 Carton et al. [11] conclude that feature-based explanations did not help users in classification tasks, Hase and Bansal [22] find them to be effective in model simulatability, which refers to the ability to predict the model behavior given an input and an explanation. Regarding example- and rule-based approaches, van der Waa et al. [57] find that both explanation types might even persuade users in cases of wrong advice.

In summary, it can be stated that the overall effect of explanations on task performance remains ambiguous. Some studies show the benefits of explanations [8], whereas others find that they can lead to humans being convinced by a wrong decision of the AI [4, 9]. Of course, this is also due to the specific setups of each study and the different goals pursued by the researchers. However, 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 start by elaborating 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. [58]. In detail, we developed a search string focusing on XAI and behavioral experiments. For both parts, 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 [51]. 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 based on the identified articles fulfilling our inclusion criteria. For each

| Source | Dataset | Datatype | Studies |
|---|---|---|---|
| Alufaisan et al. [3] | COMPAS [46] | Tabular | AI-, XAI-assisted (Anchor) |
| | Census [18] | | AI-, XAI-assisted (Anchor) |
| Bansal et al. [4] | LSAT [55] | Text | XAI-assisted (Confidence, Explain Top-1, Explain Top-2, Adaptive Expert) |
| | Book reviews [24] | | XAI-assisted (Confidence, Explain Top-1, Explain Top-2, Adaptive AI, Adaptive Expert) |
| | Beer reviews [39] | | XAI-assisted (Confidence, Explain Top-1, Explain Top-2, Adaptive AI, Adaptive Expert) |
| Buçinca et al. [8] | Fat content prediction [8] | Image | AI-,XAI-assisted (Inductive and Deductive Explanations) |
| Carton et al. [11] | Online toxicity [60] | Text | AI-, XAI-assisted (Sparse, Partial, Full explanation) |
| Lai et al. [35] | Deception detection [44] | Text | XAI-assisted (Predicted Label & Signed, Predicted Label & Guidelines, Predicted Label & Accuracy) |
| Liu et al. [38] | COMPAS [46] | Tabular | XAI-assisted (Static/Static, Interactive/Static, Interactive/Interactive) |
| | ICPSR [43] | | XAI-assisted (Static/Static, Interactive/Static, Interactive/Interactive) |
| | BIOS [16] | Text | XAI-assisted (Static/Static, Interactive/Static, Interactive/Interactive) |
| Mohseni et al. [41] | Fake news [41] | Text | XAI-assisted (Feature Importance) |
| van der Waa et al. [57] | Diabetes mellitus type 1 [57] | Tabular | AI-, XAI-assisted (Rule-based, Example-based) |
| Zhang et al. [63] | Census [18] | Tabular | AI-, XAI-assisted (Confidence, Feature Importance) |
| Fügener et al. [20] | Dog breed ImageNet [48] | Image | AI-, XAI-assisted (Suggestion and Certainty) |

**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.**

article, we extracted all individual treatments and outcomes for each study. For instance, if an experiment compared AI- with XAI-assisted decision-making 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 identified articles regarding the conducted user studies.

As empirical AI-assisted decision-making studies can vary considerably in terms of tasks, problem setting, and reported performance metrics, we further filter our set of collected studies in the following way: first, we focus on studies assessing classification tasks as they account for the largest available subset across all entries in our database. Second, we further 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. Furthermore, we exclude articles that are not peer-reviewed.

Subsequently, we extract all necessary performance metrics from the articles. We define the performance when 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. Respectively, AI assistance including explanations is called XAI assistance, and the resulting performance measure, 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' $g$ [25] to correct the SMD for a possible upward bias of the effect size when the sample size of a study is small (n≤20) as, e.g., in Zhang et al. [63]. Thus, Hedges' $g$ is smaller for n≤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 the procedure outlined in Higgins et al. [27].

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 [27].

For our meta-analytic model and the pooling of effect sizes, we estimate a random-effect 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 $\tau^2$ and its confidence intervals we use the DerSimonian-Laird estimator [17] and Jackson's method [29], respectively.

To provide further insights across current XAI studies, we conduct subgroup analyses based on experimental designs. Many studies have discussed that task choice has a strong influence on the experimental outcome [33, 34]. In this article, we focus 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. [20] 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, followed by outlining the meta-study results, which include the respective subgroup analyses.

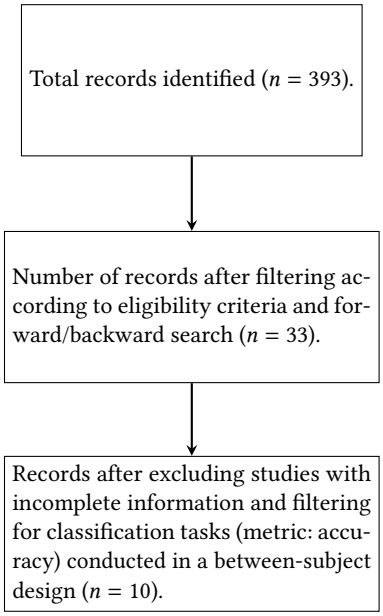


**Figure 1: Flowchart describing the data collection and article selection procedure.**

## 4.1 Data Collection Results

As of February 2022, we identified a total number of 393 articles through the iteratively developed search string. 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 10 articles in the meta-analysis and the respective subgroup analyses. Figure 1 visualizes the entire filtering process in a flowchart. Moreover, Table 1 provides an overview of all included articles together with information about each dataset, datatype, and the number of participants. 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 as displayed in Table 1.

## 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.

The results of the analysis reveal that, on average, the standardized mean difference (SMD) of all studies that reported AI- and XAI-assisted performance is 0.08 (confidence interval (CI) 95% [-0.1402, 0.2938]). A z-test against the null-hypothesis that the effect size is 0 cannot be rejected ($z = 0.69, p = 0.488$). This means we do not find a significant difference between AI-assisted and XAI-assisted performance in our current sample of studies.

Additionally, we measure an overall $I^2$ of 52.6% which is considered moderate [27]. The $\tau^2$ is 0.0616 (CI 95% [0.0105, 0.5243]) and Q is significantly different from 0 ($Q = 21.09, df = 10, p = 0.0205$). The prediction interval ranges from -0.54 to 0.69. That means, we can expect negative as well as positive effects of XAI- in comparison to AI assistance. All metrics point towards moderate heterogeneity in the analysis.

Thus, 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. [57]. The authors find an increase from 65% accuracy in the AI-assisted condition to 73% accuracy in the XAI-assisted condition. The highest negative impact of XAI is measured by Zhang et al. [63]. In the XAI condition, adding explanations decreases the AI-assisted performance by 4.5 percentage points. Figure 2 visualizes the meta-analysis.

This first result has implications for the upcoming analyses. We are limited by the current number of studies that report AI-assisted performance and related dispersion metrics. Therefore, in the following, we analyze the SMD of human and XAI-assisted performance.

## 4.3 Human vs. XAI Assistance

Next, we measure the overall effect of XAI in comparison to human performance. Therefore, we filter all studies that report human and XAI-assisted performance. This results in a sample of 34 studies and a total number of 5,163 participants. Based on this sample, we analyze whether XAI-assisted decision-making leads to performance improvements compared to humans conducting the respective tasks alone.

The results of the meta-analysis indicate that, on average, XAI assistance increases task performance by 0.59 SMD as compared to humans conducting the tasks alone. The 95% confidence interval (CI) of the SMD is 0.3855 to 0.7917. As this range does not include an effect size of zero and a z-test is significant ($z = 5.68, p < 0.0001$), 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 conditions ($Q = 219.31, df = 33, p < 0.0001$). Moreover, $I^2$ is 85% with a 95% confidence interval of 79.9% to 88.7%. The estimated $\tau^2$ is 0.2831 (CI 95% [0.1791, 0.5682]). Thus, the level of heterogeneity can be considered substantial [27]. To provide an intuitive understanding of the heterogeneity, we report also the prediction interval that represents the expected range of true effects in other studies [28]. The prediction interval in the analysis is -0.5673 to 1.7445. This means that we can not say with certainty that XAI always has a positive impact on human decision-making as the prediction interval is not exclusively larger than 0. Figure 3 visualized the forest plot of this meta-analysis.

The high heterogeneity motivates subgroup analyses to explain current differences in study results. One of the two experiments conducted by Alufaisan et al. [3] encountered the most negative effects, reducing human performance by 3.7 percentage points. The authors support humans in recidivism prediction and explained

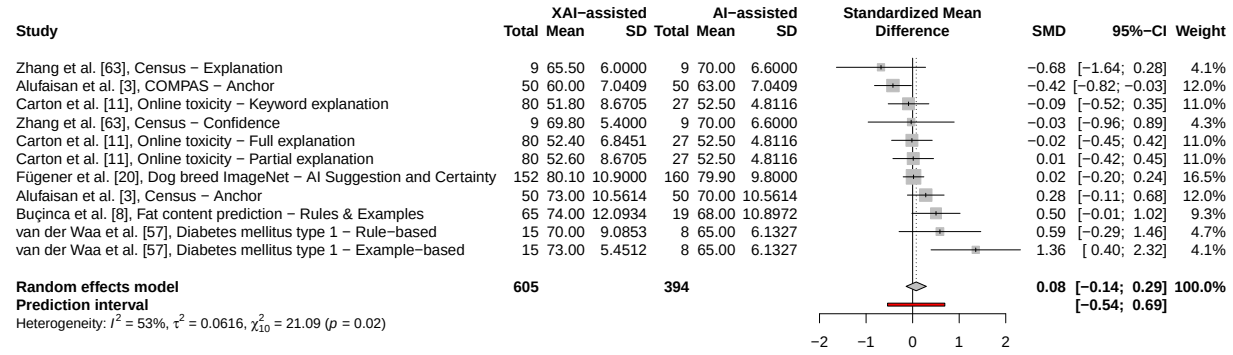


**Figure 2: AI compared to XAI-assisted performance.**

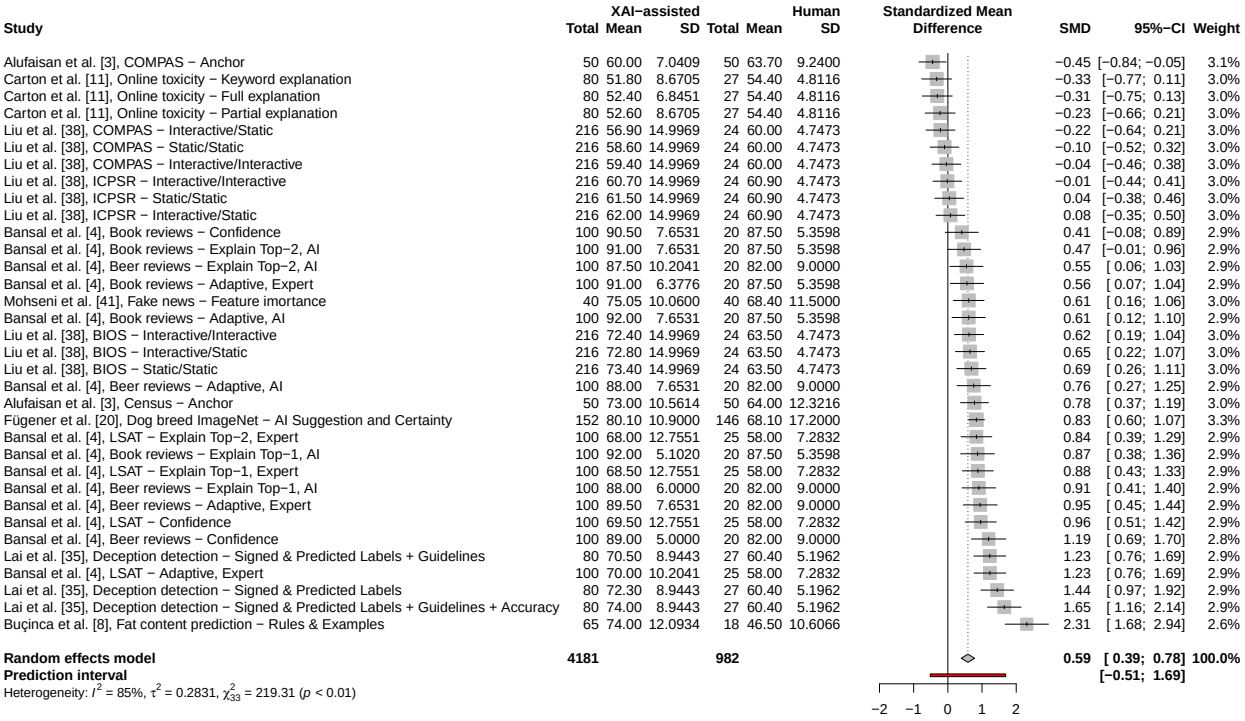


**Figure 3: Isolated human performance in comparison to XAI-assisted performance.**

the AI's reasoning based on feature importance. Due to a high dispersion, this reduction is, however, not statistically significant. The highest improvement by 27.5 percentage points can be found in an AI-assisted condition from [8]. The authors assist humans in detecting deceptive hotel reviews.

It is important to highlight that the significant SMD does not mean that including explanations will improve performance over simply providing AI advice without any form of explainability as we were not able to find a significant difference between AI-assisted and XAI-assisted performance in Section 4.2. Rather, it can be interpreted as a positive effect of some form of AI advice in comparison to human performance. In future work, we aim to also conduct additional subgroup analysis with respect to AI-assisted performance. However, in the current study, we are limited by the number of studies that reported AI-assisted dispersion metrics.

## 4.4 Tabular vs. Text Data

In addition, we conduct a subgroup analysis based on three different data types that were used in our sample—tabular, text, and image data. Only two studies report experiments using image data leading to only two image data-based treatments. Thus, the interpretation of image data is not meaningful and we excluded those studies, which

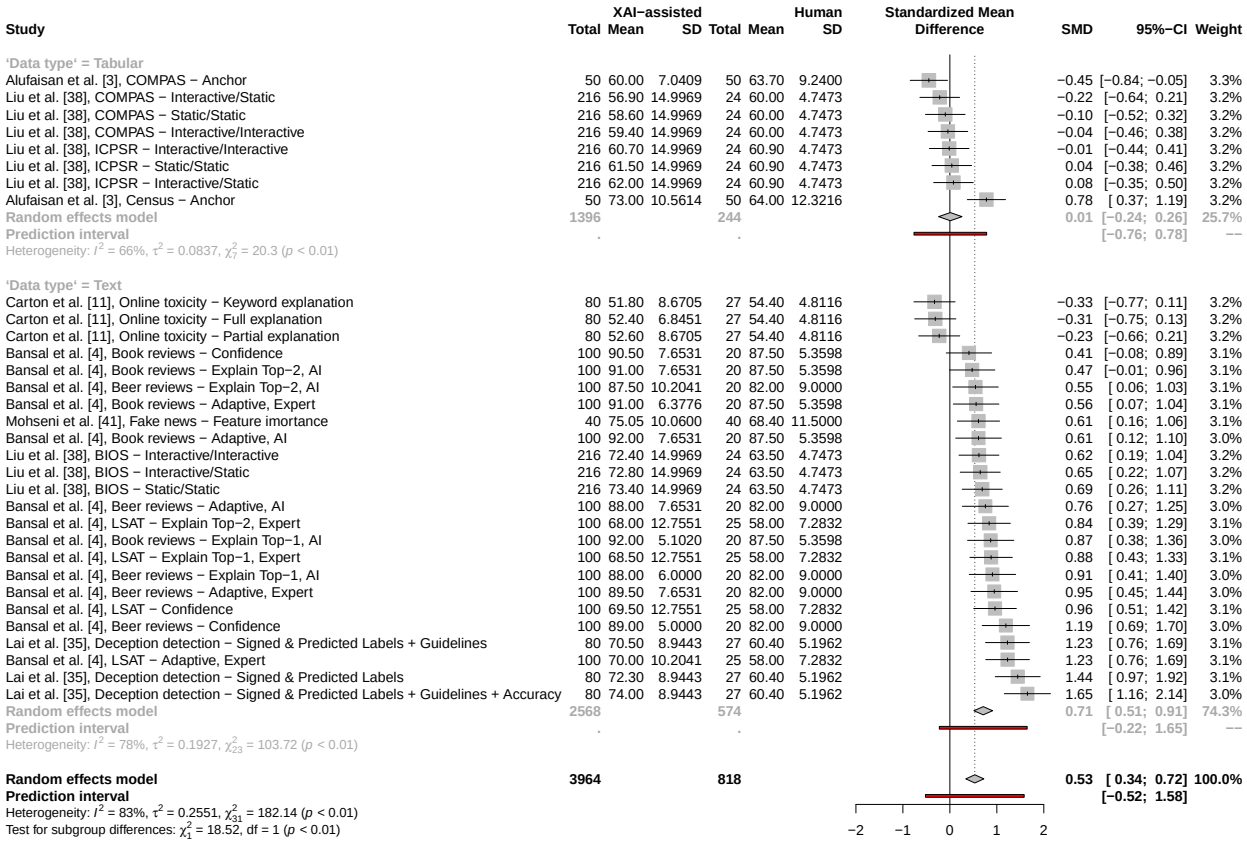


**Figure 4: Subgroup meta-analysis based on the data types of the experiments.**

leads to a sample size of 32 experiments and 4,782 participants. In the following, we first describe the tabular data subgroup and then the text data subgroup.

The tabular subgroup inhibits a SMD of 0.0101 (95% CI [-0.2388, 0.2590]). As this range does include an effect size of zero and the z-value is 0.08 with a corresponding p-value of 0.9366, we can not reject the null hypothesis. That means we cannot tell that the SMD is statistically significantly different from 0. We observe that the true effect size in this subgroup is not equal ($Q = 20.3, df = 7, p < 0.005$). The $I^2$ is 65.5% with a 95% confidence interval ranging from 26.7% to 83.8% and $\tau^2$ has a value of 0.0837 (95% CI [0.0115, 0.4721]). Finally, the prediction interval ranges from -0.7632 to 0.7834. 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 0.7108 (CI 95% [0.5110, 0.9106]). We see that this subgroup has a confidence interval larger 0 and the z-test with $z = 6.97$ shows that the SMD is significantly different from 0 ($p < 0.0001$). In comparison to the tabular data subgroup we observe a higher heterogeneity ($Q = 103.72, df = 33, p < 0.0001$). We measure a $\tau^2$ of 0.1927 (CI 95% [0.0936, 0.4282]) and an $I^2$ of 77.8% (CI 95% [67.5%; 84.9%]). The prediction interval ranges from -0.2239 to 1.6454 which means that we can impact still also some negative XAI results on text data.

Based on these two subgroups, we observe significant performance differences between tabular and text data with performance gains on text data ($Q = 18.52, df = 1, p < 0.0001$). A larger part of the overall $I^2$ (83%) can be allocated to the text data subgroup ($I^2 = 77.8$ in comparison to the lower $I^2$ in the tabular data subgroup ($I^2 = 65.5$. Furthermore, the prediction interval for tabular data is much larger than for text data.

Figure 4 displays the forest plot of the subgroup meta-analysis.

## 5 LIMITATIONS

As XAI is a rather new field of research, at least in comparison to fields where meta-analysis is more common, e.g., cancer research [19], we have some major limitations that all 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 have been recruited via online platforms such as Mechanical Turk. People conducted the task not in a controlled lab environment inducing higher variability.

Second, the studies used different XAI algorithms that ranged from just providing an additional confidence score to personalized explanations. We conduct also a subgroup analysis regarding the XAI algorithm category but are limited by the current sample size and therefore could not achieve interpretable results.

Third, also task design differed between the studies. Some studies used intuitive and easy tasks such as sentiment analysis of movies, others complicated tasks 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, 38]. Even though they have 8 studies this poses a possible limitation.

Moreover, as many studies did not report dispersion metrics numerically, we needed to extract them from 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 meta-analysis is limited with regard to drawing causal insights. Our analysis should be more seen as an overview of existing research.

The main limitation of the comparison of AI-assisted versus XAI-assisted performance is the small sample size due to many studies not reporting the dispersion of their AI-assisted condition. Based on a semi-automatic retrieval of new publications, we aim to increase successively our sample size of XAI studies. In future work, we aim to analyze a larger sample with regard to various task-specific subgroups.

## 6 DISCUSSION AND IMPLICATIONS

In this article, we conducted the first meta-analysis on XAI-assisted decision-making. Based on a structured data collection process, we collected 393 XAI-related articles. Subsequently, we applied inclusion criteria and finally were left with a set of 10 articles reporting 32 studies. We extracted performance and dispersion metrics from all of these studies and applied a statistical meta-analysis.

In our current sample, we find no significant difference between XAI- and AI-assisted performance. We observe that some studies report negative effects of XAI and others positive effects. Therefore, additional subgroup analyses should be conducted, to find study differences that explain the performance variance. However, due to the small number of studies that report dispersion metrics for AI assistance and XAI assistance, we could not conduct additional subgroup analyses to investigate differences between subgroups. While our meta-study does not currently allow for interpretation of the limitations of XAI, based on the qualitative interpretation of the current studies, we see potential improvements by conducting human-centered research on XAI.

Additionally, we find a positive effect of XAI assistance on human task performance. Since we find no difference between XAI- and AI assistance the results need to be interpreted carefully. We cannot conclude that XAI will lead to an overall performance superior to AI assistance. However, we see a tendency that some form of (X)AI assistance improves human performance. Still some articles show negative results [3, 11]. Researchers for example discuss negative potential effects due to over-reliance on AI [4, 9]. On the one hand, this means practitioners should not blindly implement explanations. On the other hand, carefully designed and implemented explanations could increase performance. Future research needs to investigate the reasons behind these performance gains. Performance increases could be due to a better understanding, increased engagement, or simply a higher acceptance of AI advice.

Furthermore, our subgroup analysis indicates a stronger positive effect of text data on performance compared with tabular data. If this effect can be isolated in future studies, it indicates that more work is needed regarding XAI assistance for tabular data. Reasons for this difference could be that text data is a more intuitive data type for humans.

Future studies, for example, could investigate the implications of certain experiment design features such as rewards or experiment length. Additionally, the absolute human or AI performance might have an impact on the relative influence of XAI. In easy tasks, for example, humans could better evaluate whether AI advice is correct or not. Furthermore, if the AI performance is very high, positive impacts could be created just by accepting AI advice on average more often. Most importantly differences between different XAI techniques should be studied.

To provide the research community with such analyses, we increase our sample size on a day-to-day basis by semi-automatically adding new relevant studies. At some point, we therefore should be able to conduct further subgroup analyses and compare AI assistance with XAI assistance on a larger sample.

## 7 CONCLUSION

In this article, we present the results from a preliminary meta-analysis of XAI-assisted decision-making. We identify a total sample of 10 articles that assess whether AI systems with and without explanations support the effectiveness of human decision-makers and report all necessary metrics needed to conduct a meta-analysis. Across these studies, we derive three major findings. First, we observe a significant positive effect of XAI assistance on user performance. Second, we do not find a significant effect of state-of-the-art explainability techniques to improve AI-assisted performance. Third, our analysis suggests that XAI assistance turns out to be particularly effective on text data compared to tabular data. These insights highlight the need for taking a human-centered perspective when developing new XAI approaches in the future.

## 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*, Vol. 35. 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. , 79–82 pages.

[6] Michael Borenstein, Larry V. Hedges, Julian P.T. Higgins, and Hannah R. Rothstein. 2010. A Basic Introduction to Fixed-Effect and Random-Effects Models for Meta-Analysis. *Research Synthesis Methods* 1, 2 (April 2010), 97–111. https://doi.org/10.1002/jrsm.12

[7] Noam Brown and Tuomas Sandholm. 2019. Superhuman AI for multiplayer poker. *Science* 365 (2019), 885 – 890.

[8] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 454–464.

[9] 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 (2021), 1–21.

[10] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019).

[11] Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2020. Feature-Based Explanations Don't Help People Detect Misclassifications of Online Toxicity. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 95–106.

[12] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. 2018. Do explanations make VQA models more predictable to a human?. In *EMNLP*.

[13] 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).

[14] Arun Das and Paul Rad. 2020. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371* (2020).

[15] Min-Yuh Day, Tun-Kung Cheng, and Jheng-Gang Li. 2018. AI robo-advisor with big data analytics for financial services. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 1027–1031.

[16] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, 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.

[17] Rebecca DerSimonian and Nan Laird. 1986. Meta-analysis in clinical trials. *Controlled clinical trials* 7, 3 (1986), 177–188.

[18] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[19] 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.

[20] 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 (MISQ)-Vol* 45 (2021).

[21] 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.

[22] 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.

[23] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), 1026–1034.

[24] 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.

[25] Larry V Hedges and Ingram Olkin. 1985. Statistical Methods for Meta-Analysis.

[26] Patrick Hemmer, Maximilian Schemmer, Michael Vössing, and Niklas Kühl. 2021. Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. In *PACIS 2021 Proceedings*.

[27] 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*.

[28] 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), e010247.

[29] 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.

[30] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna M. Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020).

[31] Jon M. Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human Decisions and Machine Predictions. *Economics of Networks eJournal* (2018).

[32] 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.

[33] Vivian Lai, Samuel Carton, and Chenhao Tan. 2020. Harnessing Explanations to Bridge AI and Humans. *arXiv* (2020). ISBN: 9781450368193 _eprint: 2003.07370.

[34] 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:2112.11471 [cs]* (Dec. 2021). http://arxiv.org/abs/2112.11471 arXiv: 2112.11471.

[35] 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.

[36] 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.

[37] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human factors* 46 1 (2004), 50–80.

[38] 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.

[39] Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining*. IEEE, 1020–1025.

[40] 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. Ledsam, 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 Shravya Shetty. 2020. International evaluation of an AI system for breast cancer screening. *Nature* 577 (2020), 89–94.

[41] Sina Mohseni, Fan Yang, Shiva Pentyala, Mengnan Du, Yi Liu, Nic Lupfer, Xia Hu, Shuiwang Ji, and Eric Ragan. 2020. Machine learning explanations to prevent overtrust in fake news detection. *arXiv* (2020). _eprint: 2007.12358.

[42] Giang Nguyen, Daeyoung Kim, and Anh Nguyen. 2021. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. *arXiv preprint arXiv:2105.14944* (2021).

[43] 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. (2014).

[44] Myle Ott, Claire Cardie, and Jeff Hancock. 2012. Estimating the prevalence of deception in online review communities. In *Proceedings of the 21st international conference on World Wide Web*. 201–210.

[45] Matteo Pennisi, Isaak Kavasidis, Concetto Spampinato, Vincenzo Schininà, 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.

[46] ProPublica. 2016. Machine Bias. (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[47] 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.

[48] 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.

[49] Max Schemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. On the Influence of Explainable AI on Automation Bias. *European Conference on Information Systems (ECIS)* (2022).

[50] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kuehl. 2022. On the Relationship Between Explanations, Fairness Perceptions, and Decisions. *arXiv preprint arXiv:2204.13156* (2022).

[51] Michiel Schotten, Wim JN Meester, Susanne Steiginga, 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.

[52] Julian Senoner, Torbjørn H. Netland, and Stefan Feuerriegel. 2021. Using Explainable Artificial Intelligence to Improve Process Quality: Evidence from Semiconductor Manufacturing. *Management Science* (2021).

[53] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, L. Sifre, Dharshan 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.

[54] 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.

[55] 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* (2017).

[56] 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*. Springer, 85–100.

[57] 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.

[58] Jan vom Brocke, Alexander Simons, Björn Niehaves, Kai Riemer, Ralf Plattfaut, and Anne Cleven. 2009. Reconstructing the giant: On the importance of rigour in documenting the literature search process. In *ECIS 2009 Proceedings*.

[59] 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, Jiyon 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 (2020), 1184 – 1194.

[60] 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.

[61] Arnold Yeung, Shalmali Joshi, Joseph Jay Williams, and Frank Rudzicz. 2020. Sequential Explanations with Mental Model-Based Policies. *arXiv preprint arXiv:2007.09028* (2020).

[62] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2019. Do I trust my machine teammate?: an investigation from perception to decision. *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019).

[63] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.