JOURNAL OF DECISION SYSTEMS
2026, VOL. 35, NO. 1, 2616693
https://doi.org/10.1080/12460125.2026.2616693

Taylor & Francis
Taylor & Francis Group

How explanations from XAI-based decision support affect human task performance: a meta-analysis

Felix Haag

Chair of Information Systems and Energy Efficient Systems, University of Bamberg, Bamberg, Germany

ABSTRACT

Artificial intelligence (AI) increasingly supports human decision-making across domains. Yet, many AI-based decision support systems (DSS) rely on machine learning models that are 'black boxes' to humans. This opacity has driven the development of explainable AI (XAI) methods that explain model outputs in human-understandable terms. Empirical findings, however, remain inconsistent regarding whether and how such explanations affect users' decision performance: some studies report improvements, while others find negligible or even negative effects. To reconcile these inconsistencies, we conduct a meta-analysis with 4589 participants comparing XAI-supported decisions to unaided decisions and 7706 participants comparing XAI-supported to AI-only supported decisions. Our analyses indicate that, on average, XAI-based decision support is associated with higher task performance compared to no support, whereas the additional gains of explanations over AI-only support are small. We also find that studies' risk of bias levels are associated with the magnitude of reported effects, suggesting larger reported performance gains in studies with higher risk of bias. Interestingly, explanation type alone does not show a significant moderating effect on task performance across studies. Overall, these findings provide a better understanding of how XAI explanations influence human decision-making and inform the design of XAI-based DSS.

ARTICLE HISTORY
Received 8 November 2025
Accepted 2 January 2026

KEYWORDS
Explainable artificial intelligence (XAI); interpretable machine learning; decision performance; explanation types; risk of bias

1. Introduction

Artificial intelligence (AI) is increasingly driving automation across a broad range of tasks, achieving levels of accuracy that often match or surpass those of human experts (see, e.g. Esteva et al., 2017; Silver et al., 2016). Yet, even as AI takes on a central role in automating tasks, responsibility for the resulting outcomes ultimately rests with human actors (Sturm et al., 2023). In response to the tension between automation and accountability, decision systems research has increasingly incorporated AI – particularly machine learning (ML) – to augment decision-making processes (Hopf et al., 2023; Mollá et al., 2024; Sekine et al., 2025).

The effective adoption of such AI-based decision support appears to depend on how well they meet user expectations. Specifically, users expect AI-based decision support systems (DSS) to be complementary, adaptive, and transparent in their outputs (Hemmer et al., 2022). Among these attributes, transparency has particularly emerged as a critical determinant of success: DSS that provide explanations for their outputs can improve users' satisfaction (Li & Gregor, 2011), trust (Wang & Benbasat, 2007), and decision performance (Gregor & Benbasat, 1999). However, a major challenge with powerful yet complex AI-based decision support lies in the reliance on ML models that are often not interpretable in human-understandable terms (Coussement et al., 2024; Kucklick, 2024). This lack of interpretability can hamper the broader adoption and effectiveness of decision support (Hemmer et al., 2022; Yeomans et al., 2019). To tackle this shortcoming, research has developed methods aimed at explaining accurate but opaque 'black box' ML model decisions (see, e.g. Lundberg et al., 2020; Mothilal et al., 2020). The tension between explainability and model performance, and the need for explainability from a user's perspective (Gregor & Benbasat, 1999), has fuelled the rise of explainable AI (XAI) (Adadi & Berrada, 2018; Bauer et al., 2021).

Given the desirable properties of explanations in DSS, XAI has gained traction as a means to support decision-making (see, e.g. Senoner et al., 2021). The interest in using explanations from XAI methods as an

CONTACT Felix Haag felix.haag@uni-bamberg.de Chair of Information Systems and Energy Efficient Systems, University of Bamberg, Kapuzinerstr. 16, Bamberg DE-96047, Germany

additional source of information has led to research exploring the collaboration of humans and XAI (Bauer et al., 2021, 2023). The use of XAI-based explanations can effectively help realise the full potential of AI in DSS, such as by minimising cognitive biases (Haag et al., 2023; Wang et al., 2019). Although XAI has attracted considerable attention, empirical evidence on its influence on task performance remains inconsistent. Certain studies demonstrate enhanced performance with XAI-supported decision-making (see, e.g. Lai et al., 2020; van der Waa et al., 2021), whereas others reveal negligible or negative effects (see, e.g. Bauer et al., 2023; Carton et al., 2020).

These inconsistent findings pose a practical dilemma for the design of DSS. Explanations in AI outputs are frequently assumed to be beneficial and are often required for reasons of transparency and accountability (Coussement et al., 2024; Meske et al., 2020). However, if explanations do not reliably improve – and may even impair – human task performance, DSS designers face uncertainty about when and how explanations should be provided.

By synthesising evidence across available empirical studies, this paper seeks to clarify how XAI-based explanations affect human performance in DSS and under which conditions they may benefit or hinder decision-making. Since most of the available studies evaluate subjects' performance when selecting among predefined decision options (see, e.g. Bauer et al., 2023; Lai et al., 2020), our analysis concentrates on classification tasks. Accordingly, our first research question (RQ) is

> RQ1: *To what extent does XAI-based decision support affect human performance in classification tasks, considering the current body of empirical studies?*

In the design of DSS, an important consideration is what defines an effective explanation that can meaningfully assist human decision-making. Building on theories from the social sciences, Miller (2019) contends that an effective – or in some sense, a 'good' – explanation is linked to the social and cognitive processes involved in transferring knowledge gained from XAI-based decision support to humans. Cognitive fit theory (CFT) posits that task performance depends on the degree of fit between task requirements and information presentation, such that an appropriate presentation can increase decision performance (Vessey, 1991). Consistent with this line of reasoning, prior research in XAI-based DSS suggests that different ways of presenting explanations may lead to different performance outcomes (Herm, 2023), although empirical findings remain overall heterogeneous (Carton et al., 2020; Haag et al., 2023; Kuhl et al., 2023).

Adopting the theoretical lens of CFT, our paper examines how different XAI explanation types – understood as alternative forms of information presentation – relate to human task performance in DSS. We thereby focus on the currently most prevalent explanation types for decision support: (i) feature attribution, (ii) counterfactual, and (iii) example-based explanations (see, e.g. Bauer et al., 2023; Herm, 2023; van der Waa et al., 2021). Hence, the second RQ is

> RQ2: *To what extent do (i) feature attribution, (ii) counterfactual, or (iii) example-based explanations affect human performance in classification tasks, considering the current body of empirical studies?*

The diversity among study setups, explanation types, but also common empirical errors (e.g. potential bias in results reporting) might account for the inconsistent findings of studies that revolve around the effect of explanations from XAI on human task performance. Meta-analyses are a powerful tool to resolve such contradictory effects in literature. Given their capabilities to account for the heterogeneity of study findings, they also allow for more reliable results than single empirical studies (King & He, 2005; Templier & Paré, 2018). Yet, current literature studies on XAI primarily focus on qualitative meta-reviews, e.g. to categorise technical methods, concepts, and related studies (Schwalbe & Finzel, 2023). To date, only one early meta-analysis conducted by Schemmer et al. (2022) quantitatively examined performance effects of XAI-based decision support across nine articles, reporting an overall positive effect. However, this analysis was necessarily limited by the small number of available studies at the time and did not examine explanation types, risks of bias in studies, or further moderators related to users' decision-making. Extending prior meta-analytic work, this paper draws on a more recent body of empirical studies, triples the number of included articles ($N = 27$), conducts differentiated sub-analyses of explanation types relevant for decision-making, and explicitly assesses the risk of bias in included studies. This allows us to estimate not only overall effects, but also whether explanation types and study quality systematically

shape performance outcomes, based on a broader and more robust empirical foundation. To address these limitations and advance understanding of how XAI-based explanations affect human task performance, we conduct a meta-analysis following three main steps. First, we perform a structured literature review to identify relevant studies within the existing research landscape. Second, we apply regression analysis to compare task performance under basic decision support (either none or AI-based decision support) versus XAI-based decision support. Third, we perform subgroup analyses to assess whether bias levels and explanation types moderate the observed effects.

Improving the understanding of the collaboration between humans and XAI is of high interest to information systems (IS) research (Bauer et al., 2021; Berente et al., 2021; Coussement et al., 2024). Whether and how contemporary explanations in AI can benefit human decision-making remains widely unclear. Accordingly, this paper presents a comprehensive and up-to-date overview of how XAI-based decision support influences task performance, along with detailed analyses of the moderating factors underlying these effects. We anticipate that our findings will offer both theoretical and practical contributions to IS and decision systems research by advancing the understanding of (a) *human-XAI collaboration* and (b) *the design of DSS that incorporate explanations obtained from XAI*.

2. Background

Starting from the definition of relevant decision support configurations and prior work on human–(X)AI collaboration in DSS, we outline key methodological properties of XAI, introduce the most prevalent explanation types for decision support, and describe cognitive fit theory as a theoretical lens for interpreting empirical findings.

2.1. Human-(X)AI collaboration in DSS

Prior DSS research and practice have focused on descriptive and diagnostic decision-aid (Power, 2008) and more recently on AI-based DSS that offer decision support of predictive and prescriptive nature (Berente et al., 2021). In particular, the latter has put forth remarkable examples of aid: DSS that complement human decision-making processes by encapsulating AI-based support have already proven effectiveness, e.g. in the operation management domain (Gonçalves et al., 2021).

Research exploring these complementary processes – emerging from the combined forces of humans and AI – is commonly described as 'human-AI collaboration' (Lai et al., 2021; Wilson & Daugherty, 2018). As this conjunction involves one or multiple human subjects as well as technical components (Seeber et al., 2020; Zschech et al., 2021), it can be described as a socio-technical process in which two or more entities (i.e. multiple humans and an AI system in this context) engage mutually in activities to accomplish common objectives (Lai et al., 2021). Human-AI collaboration is therefore a counter-concept to full task automation through AI and aims to achieve superior decision performance by compensating for the respective weaknesses (Seeber et al., 2020). For example, state-of-the-art chatbots can process large amounts of text data and provide concise summaries, while humans are able to contextualise the output through nuanced domain understanding (Daugherty & Wilson, 2018). In DSS, human-AI collaboration typically involves an individual teaming up with AI that relies on complex ML models to predict discrete classes or continuous outputs (Hemmer et al., 2022; Jussupow et al., 2021). This paper adopts this view and conceptualises AI as a collaborator in the decision-making process. Accordingly, AI-based decision support in this paper refers to predictive support, where AI provides outcome predictions intended to inform human judgement.

Due to the interest in explanations encapsulated in IS, research on human-XAI collaboration has begun to investigate the interaction between humans and XAI. Although it resembles the concept of human-AI collaboration (i.e. the parties join forces), the emphasis here lies on how explanations influence factors such as reliance and trust (Bussone et al., 2015), situational information processing, mental models (Bauer et al., 2023), and task performance (Lai & Tan, 2019). With respect to task performance, Wang et al. (2019) assume that XAI explanations can encourage people to discard personal initial hypotheses (e.g. caused by a fixation on initial decisions) and explore alternative

ones, thus, helping to avoid cognitive biases. On the contrary, Bauer et al. (2023) show that XAI explanations strongly alter the situational weighting of available information and mental models, leading to decision bias and, ultimately, to lower task performance. Despite the increasing use of XAI, this ambiguity shows that there is still great uncertainty as to whether and how XAI affects task performance.

2.2. XAI method properties and explanation types

XAI research has developed a broad range of methods (Schwalbe & Finzel, 2023), along with corresponding categorisations to distinguish between them (see, e.g. Adadi & Berrada, 2018; Arrieta et al., 2020; Meske et al., 2020) and resulting systems (see, e.g. Kucklick, 2024). We summarise three of the mentioned characteristics in the literature that are relevant to our data collection and analysis: First, XAI methods can be distinguished based on whether additional procedures are necessary to make the patterns identified by an ML model understandable to users (Adadi & Berrada, 2018). Approaches that are applied to an already trained model are known as *post-hoc methods* (e.g. permuting input data and observing corresponding changes in the model's output). In contrast, *intrinsic methods* are inherently interpretable by design (e.g. linear models) and therefore do not require extra steps to achieve explainability. Second, XAI methods differ in their flexibility regarding the types of ML models they can explain (Meske et al., 2020). *Model-specific approaches* are confined to certain algorithmic families, such as kernel-based or tree-based models. *Model-agnostic approaches*, on the other hand, can be applied to any model, making them independent of a particular model architecture or type. Third, the level of interpretation. *Global interpretable* methods focus on the model as a whole unit and attempt to explain the general logic and behaviour leading to all the outcomes. *Local interpretable* methods explain predictions at the level of a single instance (Mohseni et al., 2021). While these three properties can categorise many XAI approaches, some fall in between or cover several properties simultaneously, such as methods that provide global and local explanations (see, e.g. Lundberg et al., 2020).

The combinations of various technical properties have led to a plethora of methods and different types of how explanations are presented to users (Adadi & Berrada, 2018). Mohseni et al. (2021) summarise the most prevalent explanation types in a method and property-independent way by formulating questions a user would direct to an ML model to retrieve explanations. We outline these types, assign resulting explanation representations, and describe them accordingly (Table 1). Our study focuses on types and associated representations that are employed for decision support rather than, e.g. for developer's diagnostic purposes to evaluate ML models (see, e.g. Herm, 2023). These are explanations of the types a) '*Why?*', '*Why not?*', and '*How?*' in the form of feature-attribution and feature-importance explanations, b) '*What-else?*' represented by example-based explanations, and c) '*How-to?*' in the form of counterfactual explanations, which in our case also encompass anchor explanations – similar in interpretive intent but representing the logical opposite of counterfactuals (i.e. feature value changes that do not alter the prediction; Molnar, 2019). Hereafter, we describe these representation forms as 'explanation types'.

Building on these distinctions and focus on specific explanation types, we define XAI-based decision support as AI-driven decision support that incorporates explanations of the types included above to enhance users' understanding of model reasoning and outcomes. These explanations—whether feature-attributions,

Table 1. Outline of XAI explanation types based on Mohseni et al. (2021).

Type	Representation Form	Description
Why?	Feature-attribution/feature-importance (local and global interpretation)	Identify which input features contributed (*why*) or not contributed (*why not*) to a model's prediction.
Why not?		*Why-not* explanations enable contrastive insights by highlighting distinctions between the predicted and the anticipated outcomes. Feature-attribution methods highlight how individual inputs contribute to a specific prediction, whereas feature importance explanations indicate the relevance of features for prediction (Lundberg et al., 2018).
How?		Offer an overview of the model's internal decision logic, *how* the model works. Common visualizations include, for example, heatmaps and saliency maps (Molnar, 2019).
How-to?	Counterfactual (local interpretation)	Depict small, hypothetical changes to a model's input that would lead to a different output. These counterfactual examples illustrate what must change for an alternative prediction to occur (Mothilal et al., 2020).
What-else?	Example-based (global interpretation)	Provide comparable examples from the dataset that produce outcomes similar to the model's prediction. Specifically, *what-else* explanations use these similar examples to explain model behavior.

example-based, or counterfactual in nature—could serve to make predictive outputs more understandable to human decision-makers and to support them in *learning from AI*, that is, in deriving new insights to the problem or task at hand based on the explanatory information provided (Meske et al., 2020).

2.3. Cognitive fit theory and task-representation alignment in XAI-based DSS

Whether explanations in AI-driven DSS are effective – and whether they enable users to derive new insights about the problem at hand – seems to depend on the cognitive processes involved in transferring knowledge from an XAI-based DSS to human decision-makers (Herm, 2023; Miller, 2019).

CFT formalises this perspective by positing that task performance depends on the degree of fit between the task and information presentation in IS (Vessey, 1991). When this fit is low, users must engage in additional cognitive effort to apply the presented information to the task, which can impair task performance (Nuamah et al., 2020). Indeed, prior IS research has shown that decision performance varies systematically with the way information is presented for a specific task, with CFT providing an explanation for these differences (see, e.g. Herm, 2023; Vessey & Galletta, 1991).

In the context of XAI-based DSS, CFT offers a theoretical lens to explain the effects of explanations in AI-based decision support on human decision-making (see, e.g. Ebermann et al., 2023; Herm, 2023); Figure 1 displays the key elements of cognitive fit theory and maps these elements to XAI-based DSS. The decision task and its associated task requirements and cognitive demands define the (i) *Problem-Solving Task*, while AI-generated outputs, the explanation type, and the associated visual or symbolic representation form constitute a part of the external (ii) *Problem Representation* (Hudon et al., 2021; Vessey, 1991). Different explanation types, such as 'why' and 'how-to' explanations, present model reasoning in different representational forms, which may vary in how well they align with the cognitive requirements of the task (Herm, 2023). According to CFT, the degree of alignment between the *(i) Problem-Solving Task* and the *(ii) Problem Representation* influences the user's *(iii) Mental Representation* of the problem, that is, the internal representation through which the decision-maker understands, structures, and reasons about the task (Herm, 2023; Vessey, 1991). With increasing alignment (i.e. cognitive fit) between the task and the problem representation, users are more likely to form appropriate mental representations, facilitating an effective *(iv) Problem Solution* and ultimately achieving higher *(v) Problem-Solving Performance*, i.e. increased task performance (Shaft & Vessey, 2006). Conversely, cognitive misfit may require additional mental transformations and cognitive effort; end-users are therefore unlikely to build a representative mental representation of the task problem, which could result in reduced task performance (Herm, 2023; Vessey, 1991).

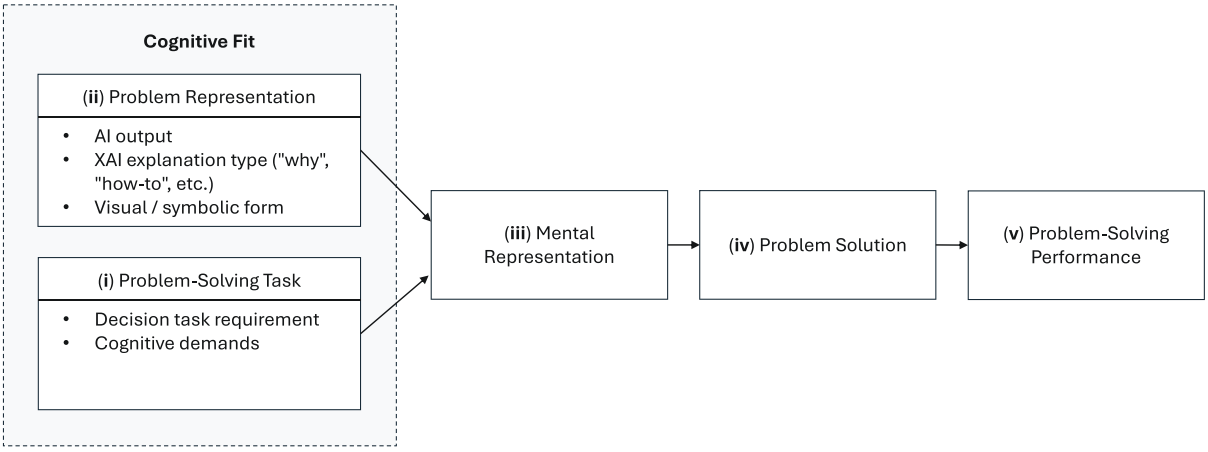

Figure 1. CFT in the context of XAI-based DSS (illustration based on Dickhaut et al. 2022; Ebermann et al. 2023; Herm 2023; Shaft and Vessey 2006).

As task requirements and explanation formats differ across decision contexts and study designs, variation in cognitive fit or misfit could help explain why XAI-based decision support improves performance in some settings but yields negligible or even negative effects in others. While CFT emphasises the importance of task-representation alignment (Vessey, 1991), a substantial share of empirical XAI studies focuses primarily on comparing explanation types (see, e.g. Nguyen et al., 2021; van der Waa et al., 2021), often without explicitly considering how the task and the respective XAI-based decision support align. This raises the question of whether explanation type alone, as part of the problem representation, is sufficient to account for performance differences, or whether the effects of explanations are contingent on task-representation alignment, as proposed by CFT.

3. Method

We start from the literature search process to obtain a final selection of eligible studies and subsequently outline the approach for statistical effect estimation and risk of bias assessment.

3.1. Literature search and data collection

When gradually developing the string for searching literature databases, we considered various aspects to cover a broad spectrum of articles. We first integrated multiple acronyms and synonyms for XAI and human task performance, as they are used interchangeably in the field (Arrieta et al., 2020). In addition to explanation types, we searched for common linked terms such as 'hypothetical' for counterfactual explanations and 'contrastive' for feature attribution explanations (Mohseni et al., 2021). We also integrated the abbreviations of the XAI methods Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) in the string, as they are among the most frequently cited approaches (Slack et al., 2020). Finally, we considered studies conducted in empirical contexts, including field and laboratory research. Our initial literature screening resulted in the following search string:

("explainable artificial intelligence" OR "xai" OR "explainable AI" OR "interpretable machine learning" OR "interpretable ml" OR "explainable machine learning" OR "explainable ml" OR ("machine learning" OR "artificial intelligence" OR "AI" AND (interpret* OR explain* OR "explanation"))) **AND**	XAI
("instance based" OR "example based" OR "counterfactual" OR "hypothetical" OR "causal" OR "anchor" OR "contrastive" OR "feature attribution" OR "feature importance" OR "LIME" OR "SHAP") **AND**	Explanation types
("task performance" OR "decision performance" OR "human accuracy" OR "human performance" OR "user study" OR "empirical study" OR "field experiment" OR "online experiment" OR "human experiment" OR "human evaluation" OR "user evaluation" OR ((behavior* OR behaviour*) AND "experiment"))	User studies

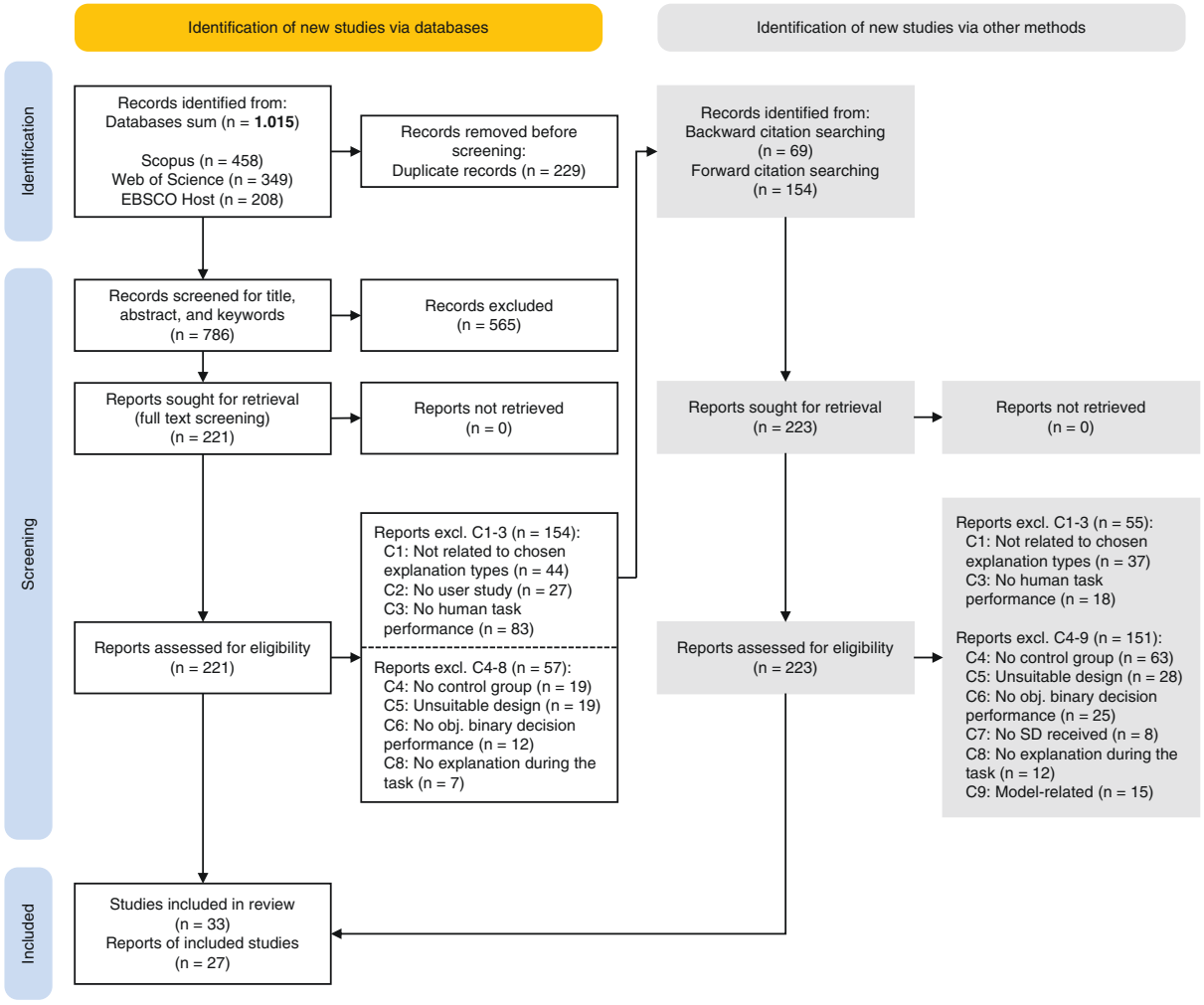

Figure 2. Literature search (PRISMA flow diagram).

The documentation of our literature search process adheres to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 standard (see Figure 2) introduced by Page et al. (2021). To capture a comprehensive range of current empirical research, we conducted searches across three academic databases: EBSCOhost, Web of Science, and Scopus ('*Identification of new studies via databases*'). To review literature in an unbiased manner, our process ensured that all screening and assessment tasks were performed independently by the author and a research assistant; all resulting deviations were resolved in joint discussion rounds.

After filtering for duplicates, the search in databases yielded a total of 786 records. We initially screened titles, abstracts, and keywords on potential relevance, resulting in 221 reports (i.e. articles) for the full-text assessment. To ensure the inclusion of a complete set of relevant reports in a subsequent forward-backward search, we divided the assessment for eligibility into two stages. In the first step, we only screened the full texts according to the criteria C1-C3 to ensure the inclusion of all thematical relevant reports for the following forward-backward search ('*Reports excluded C1-3*'). Relevant reports had to be related to at least one of the selected explanation types (i.e. feature attribution, counterfactual, or example-based explanations) (C1), describe the results of a behavioural experiment (C2), and report a quantitative measure for human task performance (C3). This assessment led to the exclusion of 154 reports. The remaining 67 reports form the basis of the forward-backward search ('*Identification of new studies via other methods*'). We identified 69 potentially relevant records from backward citation searching and 154 reports through forward citation searching relying on abstract, title, and keywords. For assessment of the resulting 223 reports, we again initially filtered reports according to criteria C1-3.

To align with the requirements of the meta-analysis, we conducted an additional curation step, refining the report sets retrieved from both the initial and forward–backward searches (i.e. 'Reports excluded C4-8' and 'Reports excluded C4-9'). This step was necessary to ensure a sufficiently low level of heterogeneity across studies (e.g. stemming from variations in research designs and tasks), the use of a common target measure, and the overall quality of the included studies (Harrer et al., 2021).

To this end, we further trimmed the set of reports according to the following criteria: To have a common ground for comparison, we could only include reports that comprise studies with a control group (i.e. no or AI-based decision support) (C4). Also, we excluded studies employing within-subject designs in which participants were exposed to multiple decision support conditions (e.g. no support and XAI-based support), as learning and carryover effects inherent to such designs render effect estimates not directly comparable to those obtained from between-subject designs (C5). Including both design types in a single meta-analysis would therefore risk biasing the pooled effect estimates, as participants' performance in later conditions may be systematically influenced by prior exposure (Harrer et al., 2021). Our analysis focuses on task performance operationalised as binary decision outcomes (e.g. correct vs. incorrect), reflecting the primary focus on classification tasks in the human-XAI literature; accordingly, we required eligible studies to either report such outcomes directly or provide sufficient information to allow a transformation into binary form (C6). We also excluded reports from which it was not possible to extract or calculate an exact standard deviation for task performance from its content or any other source (i.e. published datasets or other meta-analysis) (C7). To remedy this issue, we contacted the authors by email but received no responses from eight of them. In addition, we excluded reports in which participants did not receive XAI-based decision support during task completion but, for example, only afterwards for the next task, to minimise heterogeneity across studies (C8). Finally, we excluded 'model-related' studies in which users were asked to predict what a model would predict, thereby improving the comparability between studies (C9). In total, we identified 27 eligible reports, which are summarised in the appendix of this paper.

The data collection relied on three upfront specified decision support configurations: If a human receives a prediction from an AI system when solving a task, we refer to this as 'AI-based decision support'; when no aid is provided, we assign a 'No decision support' label. If participants receive AI-based support and are additionally provided with an explanation treatment in the AI system's output, we call such support 'XAI-based decision support'. Based on these decision support configurations, we extracted all study conditions and recorded them as separate tuples in a database. For example, if a study compared a control group with no decision support to two XAI treatments, our database documented each of the three conditions individually. For studies lacking an exact standard deviation, we used the summary statistics from the subset of articles that overlapped with those analysed by Schemmer et al. (2022). In total, our data collection resulted in 33 studies and the results of 131 individual study groups.[1]

The specified minimum number of studies required to perform valid statistical analyses varies in literature between two (see, e.g. Pigott, 2012) and ten studies (see, e.g. Higgins & Green, 2008). Given that our data collection yielded 33 studies, we assume that we obtained a suitable number to conduct meaningful analyses.

3.2. Statistical analyses and risk of bias assessment

The target variable task performance was measured as the ratio of correctly made binary decisions to the total number of decisions, providing a consistent performance indicator. For the comparisons between study conditions, we calculated the Standardized Mean Difference (SMD) using Hedges' g (Hedges, 1981), which helps prevent overestimation of the effect size in studies with a sample size of $N \leq 20$ (Higgins & Green, 2008). We anticipated that the studies included do not have a common true effect size (i.e. a fixed effect), as the experimental designs, tasks, and sample characteristics differ substantially across studies (see appendix). Therefore, instead of fixed-effects regression models, we used random-effects models to estimate the average effect size, referred to as the 'pooled effect'. To assess the significance of the pooled effect, the analyses employed the Knapp-Hartung adjustment, which reduces the likelihood of false-positive effect estimates (Hartung & Knapp, 2001). The variance of the model's estimated true effect was quantified using I^2 statistic (Higgins & Thompson, 2002), which indicates the percentage of variability in effect sizes not attributable to sampling error (Harrer et al., 2021). To control for repeated comparisons against a single group, we adjusted the sample size of each study's control group by dividing it by the number of conditions incorporated into the model (Higgins & Green, 2008). The robustness of the pooled estimates was examined using leave-one-out sensitivity analyses, in which the meta-analytic model is

recalculated after iteratively excluding one study (Harrer et al., 2021). In addition, we used standardised-residual diagnostics to identify potentially influential study conditions. The statistical analyses of this paper relied on the R packages 'meta' by Schwarzer et al. (2015) and 'dmetar' by Harrer et al. (2021).

A thorough evaluation of the included studies and the quantification of their impact on the meta-analysis results are crucial components of research synthesis. This is because the estimation of effect sizes and the testing of hypotheses in statistical analyses may be strongly driven by biases arising from study design, conduct, and reporting (Templier & Paré, 2018). Hence, this meta-analysis conducted a risk of bias analysis by employing the Cochrane Risk of Bias 2 (RoB 2) tool for randomised trials (Sterne et al., 2019). RoB 2 is divided into five risks of bias domains, where each domain contains a series of signalling questions (e.g. 'Was the allocation sequence random?') aimed at obtaining information on characteristics of studies relevant to bias. The judgement for the risk posed by a domain was calculated from an algorithm and results from the answers to the signalling questions. Ratings regarding studies' risk of bias can result in having *Low*, *Some concerns*, or *High* bias (see Sterne et al. (2019) for more details on RoB 2). Finally, we assessed the included studies for potential publication bias using Egger's regression test (Egger et al., 1997).

4. Results

We begin by comparing (a) XAI-based decision support with no decision support and (b) XAI-based decision support with AI-based support, followed by an examination of how the studies' risk of bias influences the results (RQ1). Finally, we investigate whether there is a moderating effect of the explanation type (RQ2).

4.1. Main effect of XAI-based decision support on human task performance

The analysis for RQ1 starts with examining the effect of XAI-based decision support compared to no decision support. The corresponding forest plot (Figure 3) arranges the study conditions in ascending order based on their effect sizes. In total, this analysis includes approximately 4589 participants, while the analysis on the difference between XAI- and AI-based decision support (i.e. the marginal effect of explanations in XAI-based decision support) involves 7706 participants. The influence of each condition is weighted by the inverse of its variance.

Reference/Study ID	Explanation Type/Condition	g	SE	N Treatment	N Control	SMD	95%–CI	Weight
Chen et al. (2022)/16	Feature attribution/SHAP	−0.5430	0.4463	20	7	−0.54	[−1.42; 0.33]	2.1%
Alufaisan et al. (2021)/22	Counterfactual/COMPAS	−0.4470	0.2026	50	50	−0.45	[−0.84; −0.05]	2.9%
Carton et al. (2020)/1	Feature attribution/Keyword Explanation	−0.3270	0.2237	80	27	−0.33	[−0.77; 0.11]	2.8%
Carton et al. (2020)/1	Feature attribution/Full Explanation	−0.3102	0.2236	80	27	−0.31	[−0.75; 0.13]	2.8%
Carton et al. (2020)/1	Feature attribution/Partial Explanation	−0.2264	0.2231	80	27	−0.23	[−0.66; 0.21]	2.8%
Liu et al. (2021)/4	Feature attribution/COMPAS Interactive Static	−0.2156	0.2154	216	24	−0.22	[−0.64; 0.21]	2.8%
Chen et al. (2022)/16	Feature attribution/GAM	−0.1923	0.4401	20	7	−0.19	[−1.05; 0.67]	2.2%
Silva et al. (2022)/15	Counterfactual/–	−0.1800	0.2803	39	19	−0.18	[−0.73; 0.37]	2.7%
Liu et al. (2021)/4	Feature attribution/COMPAS Static Static	−0.0974	0.2152	216	24	−0.10	[−0.52; 0.32]	2.8%
Ma et al. (2024)/24	Feature attribution/RCDV...SG	−0.0433	0.3953	32	8	−0.04	[−0.82; 0.73]	2.3%
Liu et al. (2021)/4	Feature attribution/COMPAS Interactive Interactive	−0.0417	0.2152	216	24	−0.04	[−0.46; 0.38]	2.8%
Liu et al. (2021)/4	Feature attribution/ICPSR Interactive Interactive	−0.0139	0.2152	216	24	−0.01	[−0.44; 0.41]	2.8%
Liu et al. (2021)/4	Feature attribution/ICPSR Static Static	0.0417	0.2152	216	24	0.04	[−0.38; 0.46]	2.8%
Liu et al. (2021)/4	Feature attribution/ICPSR Interactive Static	0.0765	0.2152	216	24	0.08	[−0.35; 0.50]	2.8%
Silva et al. (2022)/15	Feature importance/–	0.2943	0.2778	42	19	0.29	[−0.25; 0.84]	2.7%
Ma et al. (2024)/24	Feature attribution/RCDV...IG	0.3108	0.3997	30	8	0.31	[−0.47; 1.09]	2.3%
Ma et al. (2024)/24	Feature attribution/RCDV...SHAP	0.3286	0.3999	30	8	0.33	[−0.46; 1.11]	2.3%
Ma et al. (2024)/24	Feature attribution/RCDV...LIME	0.4174	0.4011	30	8	0.42	[−0.37; 1.20]	2.3%
Bansal et al. (2020)/12	Feature importance/Explain Top2	0.4868	0.2008	100	34	0.49	[0.09; 0.88]	2.9%
Ma et al. (2024)/25	Feature attribution/German credit...SG	0.4916	0.4023	30	8	0.49	[−0.30; 1.28]	2.3%
Bansal et al. (2020)/11	Feature importance/Explain Top2	0.5515	0.2014	100	34	0.55	[0.16; 0.95]	2.9%
Liu et al. (2021)/4	Feature attribution/BIOS Interactive Interactive	0.6191	0.2170	216	24	0.62	[0.19; 1.04]	2.8%
Bansal et al. (2020)/12	Feature importance/Adaptive	0.6259	0.2023	100	34	0.63	[0.23; 1.02]	2.9%
Liu et al. (2021)/4	Feature attribution/BIOS Interactive Static	0.6469	0.2172	216	24	0.65	[0.22; 1.07]	2.8%
Liu et al. (2021)/4	Feature attribution/BIOS Static Static	0.6887	0.2175	216	24	0.69	[0.26; 1.11]	2.8%
Bansal et al. (2020)/11	Feature importance/Adaptive	0.7447	0.2038	100	34	0.74	[0.35; 1.14]	2.9%
Alufaisan et al. (2021)/23	Counterfactual/Census	0.7783	0.2077	50	50	0.78	[0.37; 1.19]	2.9%
Chen et al. (2022)/16	Feature attribution/LIME	0.8463	0.4563	20	7	0.85	[−0.05; 1.74]	2.1%
Bansal et al. (2020)/12	Feature importance/Explain Top1	0.8658	0.2056	100	34	0.87	[0.46; 1.27]	2.9%
Bansal et al. (2020)/11	Feature importance/Explain Top1	0.8679	0.2057	100	34	0.87	[0.46; 1.27]	2.9%
Ma et al. (2024)/25	Feature attribution/German credit...LIME	0.9265	0.4116	31	8	0.93	[0.12; 1.73]	2.2%
Lai et al. (2020)/10	Feature attribution/Signed+Pred.+Guidel.	1.2262	0.2384	80	27	1.23	[0.76; 1.69]	2.8%
Lai et al. (2020)/10	Feature attribution/Signed+Pred.	1.4448	0.2442	80	27	1.44	[0.97; 1.92]	2.8%
Walter et al. (2023)/30	Feature attribution/–	1.4560	0.1627	93	99	1.46	[1.14; 1.77]	3.0%
Lai et al. (2020)/10	Feature attribution/Signed+Pred.+Guidel.+Acc.	1.6512	0.2505	80	27	1.65	[1.16; 2.14]	2.7%
Ma et al. (2024)/25	Feature attribution/German credit...IG	2.1450	0.4749	30	8	2.14	[1.21; 3.08]	2.0%
Ma et al. (2024)/25	Feature attribution/German credit...SHAP	2.3029	0.4822	31	8	2.30	[1.36; 3.25]	2.0%
Buçinca et al. (2020)/13	Example–based/–	2.3095	0.3234	65	18	2.31	[1.68; 2.94]	2.5%
Random effects model				**3667**	**922**	**0.52**	**[0.28; 0.76]**	**100.0%**
Prediction interval							**[−0.82; 1.86]**	

Heterogeneity: $I^2 = 86\%$, $\tau^2 = 0.4221$, $p < 0.01$

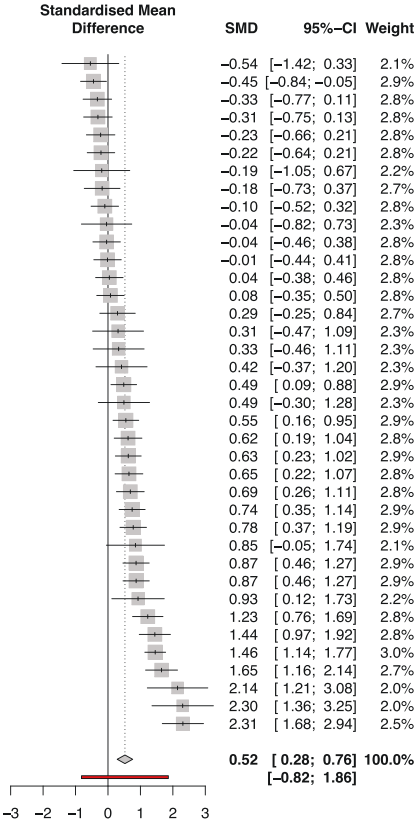

Figure 3. Forest plot on the comparison of XAI (treatment) versus no (control) decision support.

The random-effects model estimates a pooled SMD of 0.52 SMD with a 95% confidence interval (CI) of [0.28, 0.76], suggesting a moderate positive effect (Harrer et al., 2021) of XAI-based decision support relative to no decision support. The analysis reveals substantial heterogeneity across studies ($I^2 = 86\%$; $Q(37) = 265.51$, $p < 0.001$) (Higgins & Thompson, 2002), supporting the appropriateness of the random-effects model for this analysis. Testing the null hypothesis of no true effect with a *t*-test shows a statistically significant deviation of the SMD from zero ($t(37) = 4.45$, $p < 0.001$).

To further probe the heterogeneity observed in this comparison, we conduct sensitivity analyses. Leave-one-out analyses indicate that the pooled estimate remains stable across exclusions of effect-size estimates (g = 0.47–0.55), suggesting that no single study disproportionately drives the overall effect. In addition, standardised-residual diagnostics identify three influential estimates ($|z| > 2$). Excluding these estimates reduces the pooled effect to $g = 0.40$ (95% CI [0.21, 0.60]), which remains statistically significant ($p < 0.001$), while heterogeneity remains high ($I^2 = 84\%$; $Q(34) = 206.00$, $p < 0.001$). These results indicate that the observed performance advantage of XAI-based decision support over no support is not dependent on a small number of extreme study results. At the same time, the persistence of substantial heterogeneity shows that the magnitude of the effect varies across tasks and decision support implementations.

Overall, the findings for this analysis suggest that *XAI-based decision support leads to a significant improvement in human task performance for classification tasks compared to having no decision support*. It should be noted that this comparison does not disentangle the specific added value of explanations for task performance beyond AI-only support but rather reflects the overall benefit of XAI-based decision support (i.e. AI predictions with explanations) relative to having no support.

To further isolate the specific effect of explanations, the subsequent analysis compares XAI-based decision support with AI-only decision support (Figure 4). The model reports an I^2 value of 50%, indicating a moderate level of heterogeneity between studies (Higgins & Thompson, 2002). The test of heterogeneity ($Q(67) = 133.48$, $p < 0.001$) additionally supports the application of a random-effects model. The pooled SMD is 0.09 with a 95% CI of [0.01, 0.16], suggesting a small positive effect (Harrer et al., 2021) of XAI-based decision support on task performance relative to AI-only decision support. Testing the null hypothesis of no true effect with a *t*-test revealed that the pooled effect size differs significantly from zero ($t(67) = 2.39$, $p = 0.02$).

Sensitivity analyses further qualify this effect estimate. Leave-one-out analyses show that the pooled effect is robust to the exclusion of any single study, with effect sizes ranging from $g = 0.08$ to $g = 0.10$. Standardised-residual diagnostics identify five potentially influential effect estimates ($|z| > 2$). Removing these estimates yields a similar, statistically significant pooled effect ($g = 0.10$, 95% CI [0.04, 0.16], $p < 0.05$) while markedly reducing heterogeneity ($I^2 = 15\%$; $Q(62) = 73.30$, $p = 0.15$). This suggests that most between-study variability is driven by a small number of influential study conditions rather than widespread inconsistency across the studies included, likely reflecting contextual differences between studies. Importantly, the small performance advantage of XAI-based over AI-only decision support across studies remains stable and is not driven by these conditions.

In sum, these results indicate that while explanations in XAI-based decision support seem to provide a statistically significant improvement over AI-only support, their added value is small in magnitude. Thus, we conclude that *explanations in XAI-based decision support enhance human task performance in classification tasks compared to AI-only support*; however, the magnitude of this effect remains modest across the studies included.

4.2. Risk of bias assessment

To better understand the main effect of XAI-based decision support on human task performance (RQ1), we next outline the results of the risk of bias analysis. All responses to the signalling questions were cross-checked by an independent, non-authoring assistant. In summary, the algorithm rated the overall risk of bias as *Low* in seventeen studies, *Some concerns* in twelve, and *High* in four (Figure 5).

A major concern in the included studies was the risk of *bias arising from the randomization process* (D1). This is due to a lack of information on the studies' procedure for group allocation and missing tests for baseline differences between intervention groups (e.g. based on participants' characteristics). *Bias due to deviations from an intended intervention* (D2) arose from a post-randomised exclusion of eligible subjects. Another source for risk in D2 resulted from the potential interaction between

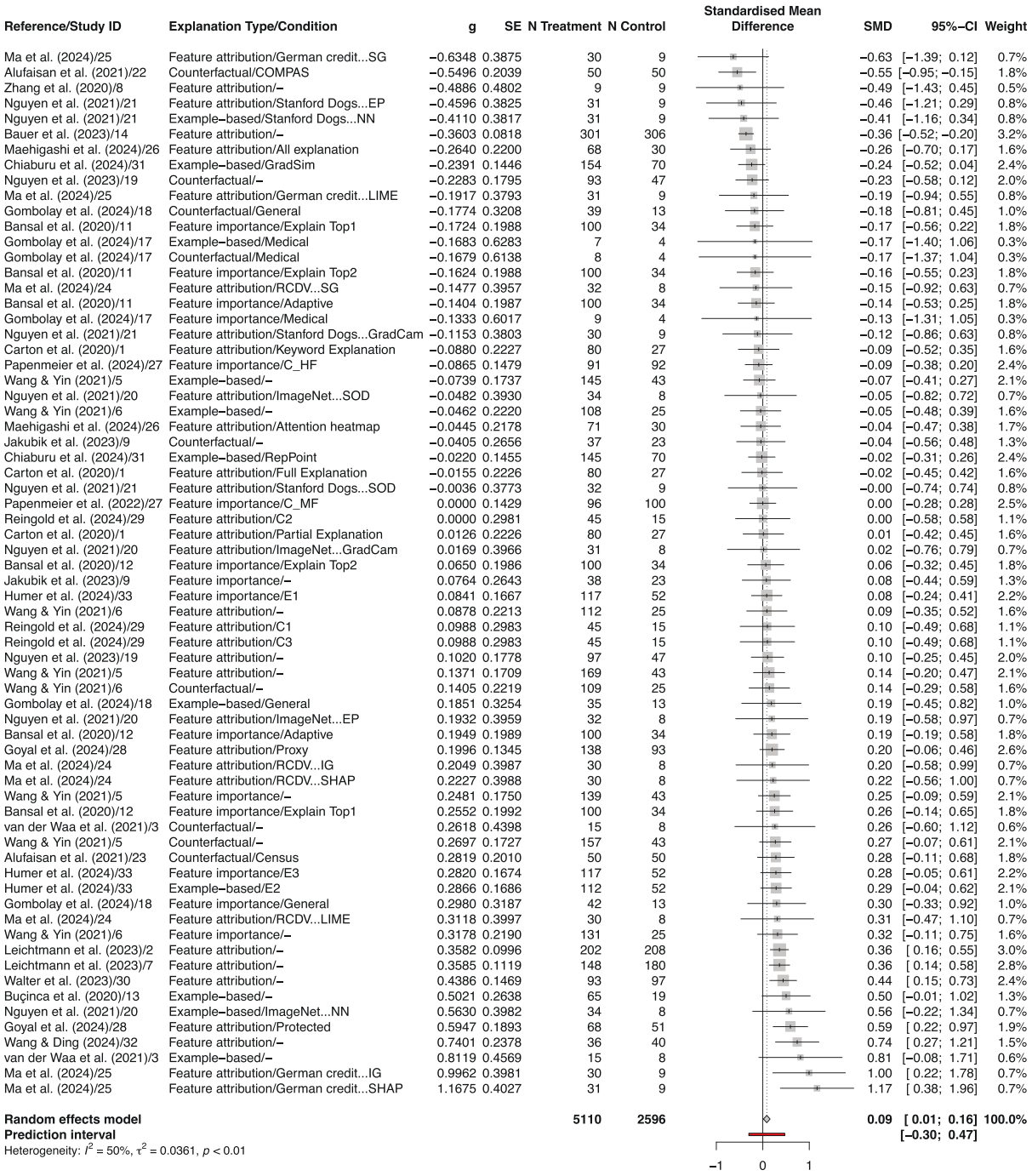

Figure 4. Forest plot on the comparison of XAI (treatment) versus AI (control) decision support.

participants before or during the experiment, which may be difficult to control due to the study setting. In general, there were no concerns about *bias due to missing outcome data* (D3) or *bias in the measurement of the outcome* (D4). A risk of *bias due to selectively reported results* (D5) primarily concerned studies that do not link a pre-specified analysis plan or approval (e.g. a study protocol) or exclude more than 10% of the participants from the analysis.[2]

The overall judgement deviated from the RoB 2 algorithm, as we assume an overall 'Low' risk for studies where concerns arose only from a missing study protocol. We made this decision because it is uncommon to pre-register studies in many research areas and for certain kinds of publications (e.g. conference proceedings). Additionally, Egger's regression tests (Egger et al., 1997) reveal no signs of publication bias in either

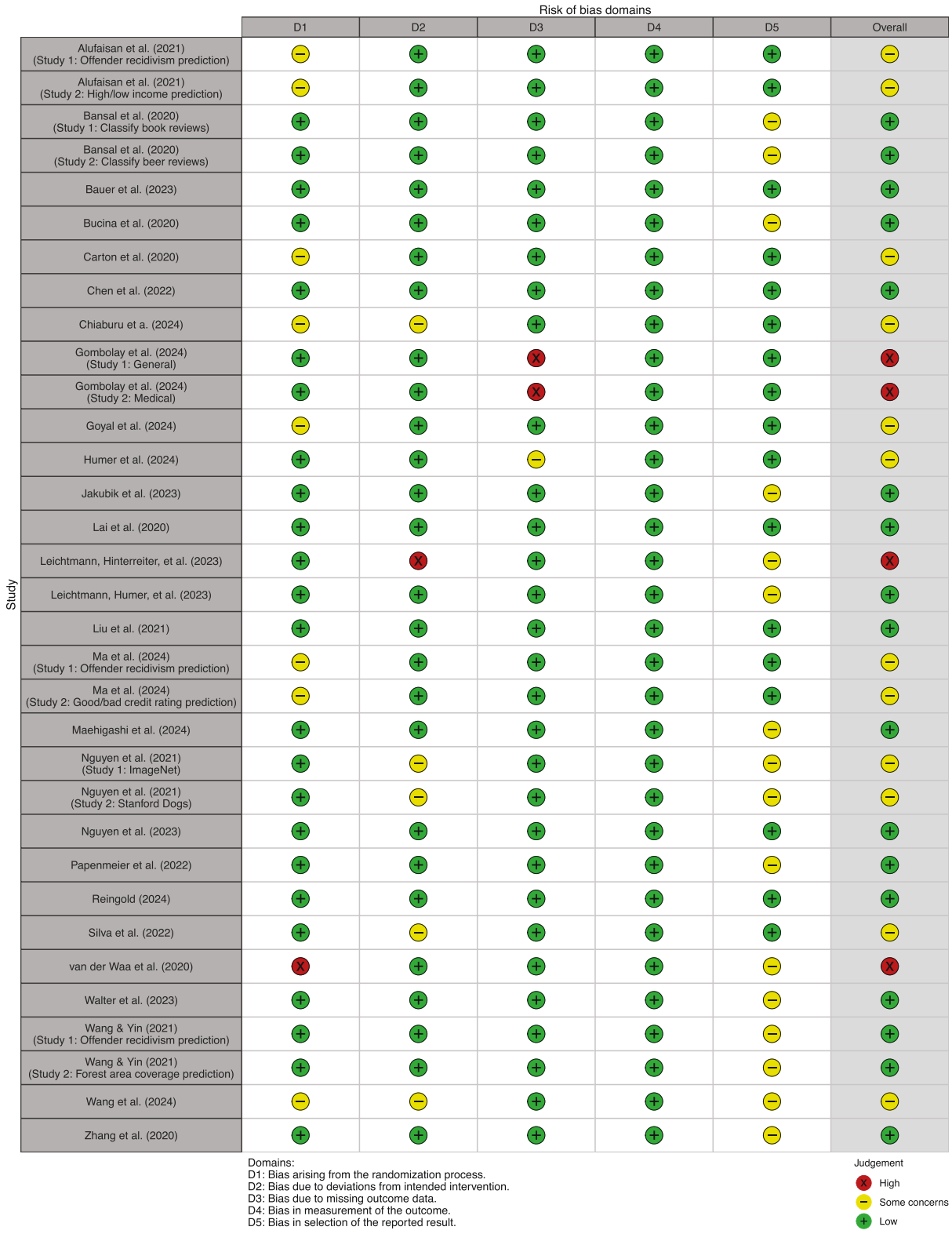

Figure 5. Risk of bias assessment using RoB 2.

comparison: for XAI-based versus no decision support, the intercept was 0.72 (95% CI [−2.44, 3.87]; $t = 0.44$, $p = 0.66$); for the case XAI-based versus AI-based decision support, the intercept was 0.27 (95% CI [−0.54, 1.07]; $t = 0.65$, $p = 0.52$). These results indicate no signs of funnel plot asymmetry, suggesting that selective publication of results is unlikely to have biased the observed effects.

Table 2. Subgroup analysis results of risk of bias assessment.

Case	XAI vs. no decision support		XAI vs. AI decision support		
Risk of Bias	Low	Some concerns	Low	Some concerns	High
SMD	0.62	0.37	0.05	0.11	0.28
[95% CI]	[0.34, 0.91]	[−0.08, 0.81]	[−0.04, 0.14]	[−0.04, 0.25]	[0.11, 0.45]
p_{type}	<0.001	<0.001	0.38	0.02	<0.01
$p_{subgroup}$	0.31			0.02	

Note: For the case "XAI vs. no decision support", we did not observe studies with an overall high risk of bias.

Building on the previously outlined comparison cases, we investigated whether there is a moderating effect of studies' risk of bias (Table 2). In the comparison of XAI-based versus no decision support, studies that have a low risk of bias label show a statistically significant positive effect (SMD = 0.62, $p_{type} < 0.001$), indicating that XAI-based decision support substantially improves task performance in these studies. In contrast, studies rated with some concerns display a smaller, statistically significant effect on the trend level (SMD = 0.37, $p_{type} = 0.06$). The test for subgroup differences does not reveal a significant moderation by risk of bias ($Q(1) = 1.03$, $p_{subgroup} = 0.31$), suggesting that the overall effect is relatively robust across bias levels. Overall, this pattern runs counter to the expectation that studies with higher risk of bias overestimate treatment effects (Harrer et al., 2021); studies with low risk of bias show a stronger effect on human task performance than those rated as having some concerns. We therefore conclude that, in this comparison, the observed performance benefits of XAI-based decision support are not affected by the studies' risk of bias.

When comparing XAI-based decision support with AI-only decision support, a clear trend emerges in which the effect size grows alongside the level of study bias. Studies assessed as having a low risk of bias reveal only a minor, statistically non-significant effect (SMD = 0.05, $p_{type} = 0.38$). In contrast, those rated as having some concerns show a slightly larger, statistically significant effect (SMD = 0.11, $p_{type} = 0.02$). Notably, studies categorised as high risk of bias demonstrate a statistically significant effect that is the largest among the categories (SMD = 0.28, $p_{type} < 0.01$). The test for subgroup differences indicates that the variations in effect sizes across bias levels are statistically significant ($Q(2) = 7.39$, $p_{subgroup} = 0.02$). These findings indicate that the level of the risk of bias seems to be systematically associated with observed outcomes with higher risk of bias linked to greater reported performance gains from explanations in XAI-based decision support. Accordingly, although XAI seems to deliver a small advantage over AI-only support (Figure 4), the bias analysis suggests that the pooled estimate may be overstated. Consistent with this pattern, excluding studies assessed as having a high risk of bias reduces the pooled effect to $g = 0.07$ (95% CI [−0.01, 0.15]), with the corresponding test for the pooled effect not reaching conventional statistical significance ($t(58) = 1.88$, $p = 0.06$), while heterogeneity remains moderate ($I^2 = 53\%$). Thus, *the genuine contribution of explanations within AI-driven decision support likely remains even smaller across the examined studies*, raising doubts about the overall benefits of current XAI-based decision support for task performance.

4.3. Effect of explanation types on human task performance

In response to the second RQ of this paper, our analysis starts by exploring how the different explanation types affect human task performance when comparing XAI-based decision support to no support. Although we identified significant differences between explanation types ($p_{subgroup} < 0.001$), the implications of these findings are somewhat limited. This is because these differences only point to variations within *certain forms of AI support* that also provide explanations for their decisions compared to no decision support. Consequently, this analysis does not provide conclusive evidence on whether the explanation type moderates the effect of XAI-based decision support.

To further explore this research question, we conduct this subgroup analysis for the case XAI-based versus AI-only decision support (Table 3). Feature attribution explanations produce a small yet statistically significant positive effect on human task performance (SMD = 0.11, $p_{type} < 0.01$). In contrast, example-based explanations (SMD = 0.07, $p_{type} = 0.62$) and counterfactual explanations (SMD = −0.02, $p_{type} = 0.83$) show no significant effect on human task performance. The subgroup analysis

Table 3. Subgroup analysis results on explanation types (XAI vs. AI based decision support).

Explanation Type	Counterfactual *How-to?*	Feature attribution *Why?/Why not?/How?*	Example-based *What-else?*
SMD	−0.02	0.11	0.07
[95% CI]	[−0.26, 0.21]	[0.02, 0.19]	[−0.14, 0.27]
p_{type}	0.83	<0.01	0.62
$p_{subgroup}$		0.49	

indicates no statistically significant differences across explanation types ($Q(2) = 1.43$, $p_{subgroup} = 0.49$), suggesting that the type of explanation does not account for variations in the observed effect of XAI-based decision support on task performance. We conclude that *the type of explanation does not systematically affect human task performance in classification tasks*.

5. Discussion

This meta-analysis led to three main findings (Table 4). We discuss these findings, outline practical and theoretical implications for research on human-XAI collaboration and DSS design, note key limitations, and finally outline future research opportunities.

5.1. Findings and implications

The first key finding (Table 4) indicates that the overall impact of XAI-based decision support on human task performance is small and varies across studies. Compared to AI-only decision support, the pooled effect size (SMD = 0.09) indicates only a minor improvement when explanations accompany AI predictions, suggesting that most performance gains stem from AI itself rather than from the explanations. This finding is consistent with Schemmer et al. (2022), who reported a similarly small effect through explanations over AI-only support that was non-significant, likely due to the limited number of nine articles included in their analysis. When XAI-based support is compared to having no decision support at all; however, the effect is considerably larger (SMD = 0.52), indicating that *some forms of AI-based decision support accompanied by explanations* can still improve performance when compared to no support. Notably, these pooled estimates were accompanied by substantial between-study heterogeneity ($I^2 = 86\%$ for XAI vs. no decision support and $I^2 = 50\%$ for XAI vs. AI-based decision support). Sensitivity analyses show that while the magnitude of the pooled effects remains stable across analyses, heterogeneity in the comparison of XAI versus AI-based decision support is largely driven by a small number of influential study conditions. This pattern indicates that, although the overall effect estimate is robust, it does not generalise uniformly across settings; rather, it reflects an amalgam of highly diverse results across task contexts and explanation instantiations. In other words, the results suggest that the impact of XAI explanations on human task performance seems highly contingent. For DSS design, these results indicate that the inclusion of XAI-based explanations cannot be assumed to improve task performance by default; instead, their effectiveness seems to depend on the conditions under

Table 4. Overview of main findings and implications.

#	Main finding	Human-XAI collaboration (theoretical implications)	Design of DSS (practical implications)
1	The overall effect of XAI-based decision support on human task performance is small and inconsistent across studies.	Reinforces that explanations in XAI-based decision systems are not universally beneficial.	Suggests that the inclusion of XAI explanations cannot be assumed to improve task performance by default.
2	Studies with higher risk of bias tend to report larger effects of XAI-based decision support on human task performance relative to AI-only support.	Indicates that variations in methodological quality and reporting practices among studies contribute to differences in reported effect sizes on human task performance.	Indicates that findings on XAI-based decision support should be interpreted in light of studies' risk of bias when deriving implications for the design of DSS.
3	The explanation type alone cannot account for the variation in reported task performance.	Shifts the focus from explanation types toward understanding how task-representation alignment and cognitive mechanisms shape the effects of XAI-based decision support on task performance.	Suggests that current evidence does not support 'one-size-fits-all' design recommendations for XAI-based DSS that rely solely on explanation type.

which explanations are provided and used – an observation consistent with CFT's emphasis on alignment between information representation and task demands.

The second finding (Table 4) concerns the subgroup analysis examining the relationship between studies' risk of bias and the reported effects of explanations on human task performance. This analysis provides a more detailed understanding of the effect estimate, revealing that studies with a higher risk of bias tend to report larger performance effects. This pattern implies that the true underlying effect of explanations – when compared to AI decision support without explanations – may be even smaller. In other words, some of the apparent performance improvement in such studies may result from methodological shortcomings rather than genuine effects of explanations. This observation underscores the importance of critically evaluating study quality when interpreting empirical findings and when translating these findings into DSS design principles. From a perspective of human-XAI collaboration, it highlights that variations in methodological rigour and reporting practices can substantially influence observed outcomes in the XAI literature. Consequently, future studies and meta-analyses should explicitly control for risk of bias to ensure that effect estimates are robust and not driven by low-quality evidence. For DSS developers, this finding highlights the need to interpret conclusions about XAI design carefully and to contextualise them based on the methodological soundness of the underlying studies.

The third finding (Table 4) revolves around the role of explanation type as a potential moderator of differences in human task performance. While prior literature has emphasised the importance of explanation design in the context of DSS and XAI-based decision support (see, e.g. Herm 2023), the current meta-analysis suggests that explanation type alone cannot explain the variability in reported effects. Specifically, the subgroup analysis on explanation types (Table 3) shows that none of the examined explanation types systematically outperforms the others in terms of task performance when compared to AI-only decision support. This indicates that the influence of explanations on task performance is shaped by a broader set of interacting factors, such as user comprehensibility and preferences (see, e.g. Wastensteiner et al., 2021), as well as characteristics of the task. Hence, these findings point to boundary conditions for the effectiveness of XAI explanations, suggesting that explanation effects cannot be understood independently.

Viewed through the lens of CFT, this pattern is plausible: user performance depends on the degree of alignment between the cognitive demands of a decision task and the way relevant information is represented (Vessey, 1991). Explanation types primarily differ in *how* model reasoning is externalised (e.g. symbolic feature weights, hypothetical alternatives, or concrete examples), but they do not uniquely determine whether this representation fits a given task. The explanation type therefore constitutes a comparatively coarse determinant of cognitive fit. In other words, the same explanation type may support cognitive fit in one task context but lead to cognitive misfit in another, depending on factors such as the nature of the task (e.g. diagnostic vs. prescriptive), the required mental operations, and the specific instantiation of the XAI explanation (e.g. visual versus textual presentation).

This interpretation is also reflected in prior empirical works. For instance, Herm (2023) shows for a medical image classification task that 'Why?'/'Why not?' explanations are associated with substantially lower reported mental effort and higher task performance compared to more information-intensive types such as 'How?' or 'How-to?' explanations that could require users to integrate broader or hypothetical content. Similarly, Bauer et al. (2023) demonstrate that 'Why?' explanations systematically reshape users' mental models and information weighting in an investment task, sometimes resulting in performance decrements, an effect that may be partly explained by explanation-induced attention that does not support the requirements of the decision task. Results from earlier studies point in a similar direction when viewed from the perspective of task demands: explanation types that encourage hypothetical reasoning, such as 'How-to?' explanations, can support exploratory or learning-oriented tasks (see, e.g. Kuhl et al., 2023), whereas in rapid classification tasks – where such reasoning is not required – 'Why?' explanations have been shown to provide little or no performance benefits (see, e.g. Carton et al., 2020).

As a result, the current findings do not support 'one-size-fits-all' design recommendations for XAI-based DSS based on explanation type alone. Instead, it suggests that effective XAI-based decision support depends on how well the decision task is aligned with the way AI outputs and explanations are represented, encompassing not only the explanation type, but also the associated visual or symbolic form through which model reasoning is externalised (see section 2.3). From a CFT perspective, this alignment matters because it shapes users' cognitive processing, such as how easily they can form task-appropriate mental

representations and integrate explanatory information into their decision-making (Vessey, 1991). Consequently, the potential benefits of explanations should be evaluated in relation to the specific task and its cognitive demands, rather than assumed a priori.

5.2. Limitations and future research

Despite our best efforts, the findings of this meta-analysis should be interpreted in light of several limitations. First, our findings are limited to specific study settings. For example, we had to exclude studies during data collection in which explanations were provided only after task completion to support learning in subsequent tasks. While these restrictions enhance comparability across studies, they may limit the scope of interpretation. Second, this meta-analysis focuses on classification tasks, which constrains the generalisability of the results to other types of decision-making, such as continuous or ranking tasks. These restrictions – both in the type of study setting and the common target variable – were necessary to ensure a common target measure and ground for comparison. In addition, the operationalisation of task performance in this meta-analysis was necessarily restricted to objective classification accuracy to ensure comparability across studies. While accuracy represents a widely used and important performance metric, it does not capture other potentially relevant outcomes of human-XAI collaboration, such as decision time, cognitive load, trust calibration, confidence, or longer-term learning effects. As a result, the present findings provide only a partial view of the impact of explanations on human behaviour and performance. Third, the scope of the subgroup analyses was necessarily limited. In our analysis, we focused on the risk of bias in studies and the type of explanation for subgroup analyses, as literature points to them as important factors that may explain variance in the target. Other potential influences, such as prediction accuracy, explanation validity, specific explanation approaches, AI literacy, and user trust, could not be included because relevant measures were not reported in most studies or exhibited high homogeneity (e.g. a dominance of post-hoc explanation approaches among the included studies). Fourth, the aggregation of different forms of AI-only decision support could have introduced additional heterogeneity. We treated AI-based decision support without explanations as a unified reference condition, although underlying implementations differed slightly across studies (e.g. prediction-only support versus prediction accompanied by a confidence score). While this aggregation was necessary to enable a common reference condition for meta-analytic comparison and sufficient statistical power, differences in baseline support may influence the estimated marginal effects of explanations. Finally, the temporal validity of our conclusions is limited. The number of studies on human-XAI collaboration is constantly increasing, leading to new insights into the mechanisms that determine its effectiveness with respect to human task performance. As the understanding of effective XAI-based decision support grows, the nature of such support (i.e. the treatment) is likely to evolve, which may in turn influence the overall effect size estimate. Hence, our results should be viewed as a snapshot of the current research landscape rather than a definitive assessment of the contribution of explanations to decision performance – a contrast to meta-analyses in other domains, such as psychology, where experimental studies (e.g. memory tasks) might tend to produce more stable effect estimates over time.

Based on the results and limitations of this meta-analysis, several directions for future research emerge. The compiled dataset could provide a foundation for such efforts and offers multiple avenues for extension. Researchers could continuously expand and update the dataset as new empirical studies on human-XAI collaboration are published. Doing so would make it possible to examine additional moderators of task performance, which could further contribute to an in-depth understanding of the collaboration of humans and XAI in DSS. Future research could also broaden the scope of performance evaluation by incorporating complementary outcome measures beyond accuracy, such as desirable learning effects or decision time, to enable a more comprehensive assessment of XAI-based decision support. In addition, future research could explore differences in task performance between decision support that relies on post-hoc explanation methods versus interpretable ML, as more user studies in the future might focus on the latter (see, e.g. Rosenberger et al., 2025). With a growing body of empirical evidence, it could also be valuable to include studies with estimation tasks on a continuous

scale (see, e.g. Haag et al., 2023). Overall, building on the current dataset and broadening its analytical scope will enable future research to further refine theoretical understanding of human-XAI collaborations, and guide the design of DSS to more effectively integrate human and XAI capabilities.

6. Conclusion

This meta-analysis synthesises available empirical evidence on how explanations from XAI-based decision support affect human task performance in classification tasks. The results indicate that while AI-based decision support can enhance human task performance, explanations do not consistently translate into additional performance gains beyond AI predictions alone. Reported positive effects tend to be associated with studies at higher risk of bias, while the type of explanation alone does not explain systematic differences across studies. By clarifying these relationships, this analysis contributes to a better understanding of how explanations can support human decision-making and highlights the importance of alignment between task demands and information representation, as emphasised by CFT (Vessey, 1991). From a practical perspective, the findings suggest that explanations should not be assumed to be universally beneficial in XAI-based DSS. In sum, this meta-analysis advances theoretical discussions on human-XAI collaboration and outlines constraints for DSS design.

Notes

1. To facilitate the reproducibility of the results, we will provide the repository containing the dataset used in this meta-analysis upon request.
2. Upon request, we share further details on the answers to the signalling question and the risk of bias judgement.

Acknowledgments

I am grateful to Carlo Stingl and Nasima Fakir for their assistance with literature checks, and to Thorsten Staake and Gerit Wagner for their helpful comments on an earlier version of this manuscript. Portions of this manuscript were edited for language clarity using ChatGPT (OpenAI, GPT-5, 2025). The author reviewed and verified all AI-assisted content.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The author(s) reported there is no funding associated with the work featured in this article.

ORCID

Felix Haag http://orcid.org/0009-0005-2227-2490

Data availability statement

This paper uses publicly available data described and referenced in the manuscript.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Alufaisan, Y., Marusich, L.R., Bakdash, J.Z., Zhou, Y., & Kantarcioglu, M. (2021). Does explainable artificial intelligence improve human decision-making? *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(8), 6618–6626. https://doi.org/10.1609/aaai.v35i8.16819

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M.T., & Weld, D. (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* (pp. 1–16). Association for Computing Machinery (ACM). https://doi.org/10.1145/3411764.3445717

Bauer, K., Hinz, O., van der Aalst, W., & Weinhardt, C. (2021). Explain it to me - explainable AI and information systems research. *Business & Information Systems Engineering*, *63*(2), 79–82. https://doi.org/10.1007/s12599-021-00683-2

Bauer, K., Von Zahn, M., & Hinz, O. (2023). Explained: The impact of explainable artificial intelligence on users' information processing. *Information Systems Research*, *34*(4), 1582–1602. https://doi.org/10.1287/isre.2023.1199

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, *45*(3), 1433–1450. https://doi.org/10.25300/MISQ/2021/16274

Buçinca, Z., Lin, P., Gajos, K.Z., & Glassman, E.L. (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* (pp. 454–464). Association for Computing Machinery (ACM). https://doi.org/10.1145/3377325.3377498

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics* (pp. 160–169). IEEE.

Carton, S., Mei, Q., & Resnick, P. (2020). Feature-based explanations don't help people detect misclassifications of online toxicity. In *Proceedings of the 14th International AAAI conference on web and social media* (pp. 95–106). AAAI Press.

Chen, V., Johnson, N., Topin, N., Plumb, G., & Talwalkar, A. (2022). Use-case-grounded simulations for explanation evaluation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems* (Vol. 35, pp. 1764–1775). Curran Associates, Inc.

Chiaburu, T., Haußer, F., & Bießmann, F. (2024). *Confident teacher, confident student? A novel user study design for investigating the didactic potential of explanations and their impact on uncertainty* (No. arXiv: 2409.17157). *arXiv*. http://arxiv.org/abs/2409.17157

Coussement, K., Abedin, M.Z., Kraus, M., Maldonado, S., & Topuz, K. (2024). Explainable AI for enhanced decision-making. *Decision Support Systems*, *184*, 114276. https://doi.org/10.1016/j.dss.2024.114276

Daugherty, P.R., & Wilson, H.J. (2018). *Human + machine: Reimagining work in the age of AI*. Harvard Business Press.

Dickhaut, E., Li, M.M., Janson, A., & Leimeister, J.M. (2022). The role of design patterns in the development and legal assessment of lawful technologies. *Electronic Markets*, *32*(4), 2311–2331. https://doi.org/10.1007/s12525-022-00597-1

Ebermann, C., Selisky, M., & Weibelzahl, S. (2023). Explainable AI: The effect of contradictory decisions and explanations on users' acceptance of AI systems. *International Journal of Human-Computer Interaction*, *39*(9), 1807–1826. https://doi.org/10.1080/10447318.2022.2126812

Egger, M., Smith, G.D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ (Clinical Research Ed)*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), Article 7639. 115–118. https://doi.org/10.1038/nature21056

Gombolay, G.Y., Silva, A., Schrum, M., Gopalan, N., Hallman-Cooper, J., Dutt, M., & Gombolay, M. (2024). Effects of explainable artificial intelligence in neurology decision support. *Annals of Clinical and Translational Neurology*, *11*(5), 1224–1235. https://doi.org/10.1002/acn3.52036

Gonçalves, J.N.C., Cortez, P., Carvalho, M.S., & Frazão, N.M. (2021). A multivariate approach for multi-step demand forecasting in assembly industries: Empirical evidence from an automotive supply chain. *Decision Support Systems*, *142*, 113452. https://doi.org/10.1016/j.dss.2020.113452

Goyal, N., Baumler, C., Nguyen, T., & Daumé Iii, H. (2024). The impact of explanations on fairness in human-AI decision-making: Protected vs proxy features. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (pp. 155–180). Association for Computing Machinery (ACM). https://doi.org/10.1145/3640543.3645210

Gregor, S., & Benbasat, I. (1999). Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly*, *23*(4), 497–530. https://doi.org/10.2307/249487

Haag, F., Stingl, C., Zerfass, K., Hopf, K., & Staake, T. (2023). Overcoming anchoring bias: The potential of AI and XAI-based decision support. In *Proceedings of the 44th International Conference on Information Systems*. Association for Information Systems (AIS).

Harrer, M., Cuijpers, P., Furukawa, T.A., & Ebert, D.D. (2021). *Doing meta-analysis with R: A hands-on guide*. CRC press.

Hartung, J., & Knapp, G. (2001). A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine*, *20*(24), 3875–3889. https://doi.org/10.1002/sim.1009

Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128. https://doi.org/10.3102/10769986006002107

Hemmer, P., Schemmer, M., Riefle, L., Rosellen, N., Vössing, M., & Kühl, N. (2022). Factors that influence the adoption of human-AI collaboration in clinical decision-making. In *Proceedings of the 30th European Conference on Information Systems*. Association for Information Systems (AIS).

Herm, L.-V. (2023). Impact of explainable AI on cognitive load: Insights from an empirical study. In *Proceedings of the 31st European Conference on Information Systems*. Association for Information Systems (AIS).

Higgins, J.P., & Green, S. (2008). *Cochrane handbook for systematic reviews of interventions*.
Higgins, J.P., & Thompson, S.G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. https://doi.org/10.1002/sim.1186
Hopf, K., Weigert, A., & Staake, T. (2023). Value creation from analytics with limited data: A case study on the retailing of durable consumer goods. *Journal of Decision Systems*, *32*(2), 289–325. https://doi.org/10.1080/12460125.2022.2059172
Hudon, A., Demazure, T., Karran, A., Léger, P.-M., & Sénécal, S. (2021). Explainable artificial intelligence (XAI): How the visualization of AI predictions affects user cognitive load and confidence. In F. D. Davis, R. Riedl, J. VOM. Brocke, P.-M. Léger, A. B. Randolph, & G. Müller-Putz (Eds.), *Information systems and neuroscience* (Vol. 52, pp. 237–246). Springer International Publishing. https://doi.org/10.1007/978-3-030-88900-5_27
Humer, C., Hinterreiter, A., Leichtmann, B., Mara, M., & Streit, M. (2024). Reassuring, misleading, debunking: Comparing effects of XAI methods on human decisions. *ACM Transactions on Interactive Intelligent Systems*, *14*(3), 1–36. https://doi.org/10.1145/3665647
Jakubik, J., Schöffer, J., Hoge, V., Vössing, M., & Kühl, N. (2022). An empirical evaluation of predicted outcomes as explanations in human-AI decision-making. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 353–368). Springer.
Jussupow, E., Spohrer, K., Heinzl, A., & Gawlitza, J. (2021). Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research*, *32*(3), 713–735. https://doi.org/10.1287/isre.2020.0980
King, W.R., & He, J. (2005). Understanding the role and methods of meta-analysis in IS research. *Communications of the Association for Information Systems*, *16*. https://doi.org/10.17705/1CAIS.01632
Kucklick, J.-P. (2024). HIEF: A holistic interpretability and explainability framework. *Journal of Decision Systems*, *33*(3), 335–375. https://doi.org/10.1080/12460125.2023.2207268
Kuhl, U., Artelt, A., & Hammer, B. (2023). Let's go to the alien zoo: Introducing an experimental framework to study usability of counterfactual explanations for machine learning. *Frontiers in Computer Science*, *5*, 1087929. https://doi.org/10.3389/fcomp.2023.1087929
Lai, V., Liu, H., & Tan, C. (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* (pp. 1–13). Association for Computing Machinery (ACM). https://doi.org/10.1145/3313831.3376873
Lai, V., & Tan, C. (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* (pp. 29–38). Association for Computing Machinery (ACM). https://doi.org/10.1145/3287560.3287590
Lai, Y., Kankanhalli, A., & Ong, D. (2021). Human-AI collaboration in healthcare: A review and research agenda. In *Proceedings of the 54th Hawaii International Conference on System Sciences* (pp. 390–399). University of Hawaii.
Leichtmann, B., Hinterreiter, A., Humer, C., Streit, M., & Mara, M. (2023). Explainable artificial intelligence improves human decision-making: Results from a mushroom picking experiment at a public art festival. *International Journal of Human-Computer Interaction*, 1–18. https://doi.org/10.1080/10447318.2023.2221605
Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., & Mara, M. (2023). Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior*, *139*, 139. https://doi.org/10.1016/j.chb.2022.107539
Li, M., & Gregor, S. (2011). Outcomes of effective explanations: Empowering citizens through online advice. *Decision Support Systems*, *52*(1), 119–132. https://doi.org/10.1016/j.dss.2011.06.001
Liu, H., Lai, V., & Tan, C. (2021). Understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–45. https://doi.org/10.1145/3479552
Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9
Lundberg, S.M., Erion, G.G., & Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *ArXiv Preprint* arXiv: 1802.03888. https://arxiv.org/abs/1802.03888
Ma, J., Lai, V., Zhang, Y., Chen, C., Hamilton, P., Ljubenkov, D., Lakkaraju, H., & Tan, C. (2024). *OpenHEXAI: An open-source framework for human-centered evaluation of explainable machine learning* (no. arXiv: 2403.05565). ArXiv. http://arxiv.org/abs/2403.05565
Maehigashi, A., Fukuchi, Y., & Yamada, S. (2024). Adjusting amount of AI explanation for visual tasks. In *Extended abstracts of the CHI conference on human factors in computing systems* (pp. 1–7). Association for Computing Machinery (ACM). https://doi.org/10.1145/3613905.3650802
Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2020). Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities. *Information Systems Management*, *39*(1), 1–11. https://doi.org/10.1080/10580530.2020.1849465
Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mohseni, S., Zarei, N., & Ragan, E.D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems*, *11*(3–4), 1–45. https://doi.org/10.1145/3387166

Mollá, N., Bossler, A., & Rabasa, A. (2024). Data stream solution for decision-making processes: A general and adaptive system for decision support. *Journal of Decision Systems*, *33*(sup1), 337–348. https://doi.org/10.1080/12460125.2024.2354590

Molnar, C. (2019). Interpretable machine learning. https://christophm.github.io/interpretable-ml-book/

Mothilal, R.K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 607–617). Association for Computing Machinery (ACM). https://doi.org/10.1145/3351095.3372850

Nguyen, G., Kim, D., & Nguyen, A. (2021). The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. *Advances in Neural Information Processing Systems*, *34*, 26422–26436.

Nguyen, T., Xu, J., Roy, A., Daumé, H., & Carpuat, M. (2023). Towards conceptualization of "fair explanation": Disparate impacts of anti-Asian hate speech explanations on content moderators. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 9696–9717). Scopus.

Nuamah, J.K., Seong, Y., Jiang, S., Park, E., & Mountjoy, D. (2020). Evaluating effectiveness of information visualizations using cognitive fit theory: A neuroergonomics approach. *Applied Ergonomics*, *88*, 103173. https://doi.org/10.1016/j.apergo.2020.103173

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E. McDonald, S. … Whiting, P. (2021). The PRISMA, 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *71*, n71. https://doi.org/10.1136/bmj.n71

Papenmeier, A., Kern, D., Englebienne, G., & Seifert, C. (2022). It's complicated: The relationship between user trust, model accuracy and explanations in AI. *ACM Transactions on Computer-Human Interaction*, *29*(4), 1–33. https://doi.org/10.1145/3495013

Pigott, T.D. (2012). *Advances in meta-analysis*. Springer US. https://doi.org/10.1007/978-1-4614-2278-5

Power, D.J. (2008). Understanding data-driven decision support systems. *Information Systems Management*, *25*(2), 149–154. https://doi.org/10.1080/10580530801941124

Reingold, O., Shen, J.H., & Talati, A. (2024). Dissenting explanations: Leveraging disagreement to reduce model overreliance. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(19), 21537–21544. https://doi.org/10.1609/aaai.v38i19.30151

Rosenberger, J., Schröppel, P., Kruschel, S., Kraus, M., Zschech, P., & Förster, M. (2025). Navigating the Rashomon effect: How personalization can help adjust interpretable machine learning models to individual users. In *Proceedings of the 33rd European Conference on Information Systems*. Association for Information Systems (AIS).

Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., & Vössing, M. (2022). A meta-analysis of the utility of explainable artificial intelligence in human-AI decision-making. In *Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society* (pp. 617–626). Association for Computing Machinery (ACM). https://doi.org/10.1145/3514094.3534128

Schwalbe, G., & Finzel, B. (2023). A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, *38*(5), 3043–3101. https://doi.org/10.1007/s10618-022-00867-8

Schwarzer, G., Carpenter, J.R., & Rücker, G. (2015). *Meta-analysis with R*. Springer International Publishing. https://doi.org/10.1007/978-3-319-21416-0

Seeber, I., Bittner, E., Briggs, R.O., De Vreede, T., De Vreede, G.-J., Elkins, A., Maier, R., Merz, A.B., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information & Management*, *57*(2), 103174. https://doi.org/10.1016/j.im.2019.103174

Sekine, Y., Kasuya, S., & Tago, K. (2025). Improving emotion estimation through a combination of ChatGPT and deep learning. *Journal of Decision Systems*, *34*(1), 1–18. https://doi.org/10.1080/12460125.2024.2440024

Senoner, J., Netland, T., & Feuerriegel, S. (2021). *Using explainable artificial intelligence to improve process quality: Evidence from semiconductor manufacturing*. Management Science.

Shaft, T.M., & Vessey, I. (2006). The role of cognitive fit in the relationship between software comprehension and modification. *MIS Quarterly*, *30*(1), 29–55. https://doi.org/10.2307/25148716

Silva, A., Schrum, M., Hedlund-Botti, E., Gopalan, N., & Gombolay, M. (2022). Explainable artificial intelligence: Evaluating the objective and subjective impacts of xAI on human-agent interaction. *International Journal of Human-Computer Interaction*, *39*(7), 1–15. https://doi.org/10.1080/10447318.2022.2101698

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), Article 7587. 484–489. https://doi.org/10.1038/nature16961

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society* (pp. 180–186). Association for Computing Machinery (ACM). https://doi.org/10.1145/3375627.3375830

Sterne, J.A.C., Savović, J., Page, M.J., Elbers, R.G., Blencowe, N.S., Boutron, I., Cates, C.J., Cheng, H.-Y., Corbett, M.S., Eldridge, S.M., Emberson, J.R., Hernán, M.A., Hopewell, S., Hróbjartsson, A., Junqueira, D.R., Jüni, P., Kirkham, J.J., Lasserson, T., Li, T., & Higgins, J.P.T. (2019). Rob 2: A revised tool for assessing risk of bias in randomised trials. *BMJ*, l4898. https://doi.org/10.1136/bmj.l4898

Sturm, T., Pumplun, L., Gerlach, J.P., Kowalczyk, M., & Buxmann, P. (2023). Machine learning advice in managerial decision-making: The overlooked role of decision makers' advice utilization. *The Journal of Strategic Information Systems*, *32*(4), 101790. https://doi.org/10.1016/j.jsis.2023.101790

Templier, M., & Paré, G. (2018). Transparency in literature reviews: An assessment of reporting practices across review types and genres in top IS journals. *European Journal of Information Systems*, *27*(5), 503–550. https://doi.org/10.1080/0960085X.2017.1398880

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, *291*, 291. https://doi.org/10.1016/j.artint.2020.103404

Vessey, I. (1991). Cognitive fit: A theory-based analysis of the graphs versus tables literature. *Decision Sciences*, *22*(2), 219–240. https://doi.org/10.1111/j.1540-5915.1991.tb00344.x

Vessey, I., & Galletta, D. (1991). Cognitive fit: An empirical study of information acquisition. *Information Systems Research*, *2*(1), 63–84. https://doi.org/10.1287/isre.2.1.63

Walter, M.C., Broder, H.R., & Förster, M. (2023). Boosting benefits, offsetting obstacles - the impact of explanations on AI users' task performance. In *Proceedings of the International Conference on Wirtschaftsinformatik*. Association for Information Systems (AIS).

Wang, D., Yang, Q., Abdul, A., & Lim, B.Y. (2019). Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–15). Association for Computing Machinery (ACM). https://doi.org/10.1145/3290605.3300831

Wang, P., & Ding, H. (2024). The rationality of explanation or human capacity? Understanding the impact of explainable artificial intelligence on human-AI trust and decision performance. *Information Processing & Management*, *61*(4), 103732. https://doi.org/10.1016/j.ipm.2024.103732

Wang, W., & Benbasat, I. (2007). Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems*, *23*(4), 217–246. https://doi.org/10.2753/MIS0742-1222230410

Wang, X., & Yin, M. (2021). Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (pp. 318–328). Association for Computing Machinery (ACM). https://doi.org/10.1145/3397481.3450650

Wastensteiner, J., Weiss, T., Haag, F., & Hopf, K. (2021). Explainable AI for tailored electricity consumption feedback-an experimental evaluation of visualizations. In *Proceedings of the 29th European Conference on Information Systems*.

Wilson, H.J., & Daugherty, P.R. (2018). Collaborative intelligence: Humans and AI are joining forces. *Harvard Business Review*, *96*(4), 114–123.

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403–414. https://doi.org/10.1002/bdm.2118

Zhang, Y., Liao, Q.V., & Bellamy, R.K.E. (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* (pp. 295–305). Association for Computing Machinery (ACM). https://doi.org/10.1145/3351095.3372852

Zschech, P., Walk, J., Heinrich, K., Vössing, M., & Kühl, N. (2021). A picture is worth a collaboration: Accumulating design knowledge for computer-vision-based hybrid intelligence systems. In *Proceedings of the 29th European Conference on Information Systems*.

Appendix Report summary

Study ID(s)	Report	Task	Description	Source/Origin	Control conditions	Explanation type(s)
22 23	Alufaisan et al. (2021)	Predict income (Census) and whether defendants will recidivate (COMPAS)	The report conducts two studies and employs explanations based on anchor LIME. XAI thereby provides explanations in the form of decision rules to explain the AI's output. The results showed that while AI predictions improved decision accuracy compared to no support, the explanations did not significantly enhance participants' task performance. Participants thereby tended to rely on AI predictions more frequently when they are correct; interestingly, XAI did not change this behaviour. The authors assume that information overload might play a crucial role for the minimal impact of XAI on performance.	Forward Search	No support AI	How-to? Method: CF/ Anchor
11 12	Bansal et al. (2021)	Sentiment analysis of beer and book reviews as well as law school questions	The report investigates within three studies the impact of (X)AI-based support in text-based tasks. Treatment group participants receive either AI-based support in the form of predictions with confidence ratings or various forms of FI explanations (i.e. word highlighting) provided by experts or XAI based on LIME. We thereby only consider XAI-based explanations. The control group ('human') answered tasks without any support. The authors found that overall XAI contributed to improve task performance compared to scenarios no or AI-only support. Participants were generally more inclined to accept AI advice when explanations were provided, regardless of the explanations' accuracy.	Backward Search	No support AI	Why? Method: FI
14	Bauer et al. (2023)	Decide whether a borrower will make a repayment	Participants play one-shot investment games in which they can invest in borrowers. The AI-based support offered predictions for repayment of an initial investment, where XAI-based support displayed attributions of borrower's characteristics based on LIME. Results indicated that explanations influenced participants' mental models and their weighting of information. This sometimes led to over-reliance on incorrect predictions and thus to lower task performance when compared to AI-only support.	Forward Search	AI	Why? Method: FA
13	Buçinca et al. (2020)	Estimate if the fat content of a meal is higher than a threshold	The report comprises a study where participants had to guess model decisions and a real-world task where participants are presented with several example images of meals that contain a certain proportion of fat. Participants then decide whether they believe that a new meal has as much fat (e.g. 30%) or more with no support, AI, or XAI-based support. XAI relies on example-based and deductive explanations. For this meta-analysis, we focus on the actual decision-making task and example-based explanations only. The results show that example-based explanations enhance users' ability to detect prediction errors. Also, participants receiving XAI support outperformed those relying on standard AI assistance, who in turn achieved better task performance than unsupported users.	Backward Search	No support AI	What-else? Method: EB

(*Continued*)

(Continued).

Study ID(s)	Report	Task	Description	Source/Origin	Control conditions	Explanation type(s)
1	Carton et al. (2020)	Decide whether an online comment is toxic	Within this study, the authors employ the task of predicting toxicity of social media comments. The ground truth stems from a pre-labelled dataset. Participants receive either no support, AI predictions or XAI support. For XAI, they analyse three forms of explanations: full explanations highlight entire text to identify toxic content. Partial explanations only points to the passage identified as most toxic, while keyword-based explanations only highlight individual toxic words. The study found that AI support either with or without explanations resulted in a slightly decreasing task performance, compared to no support. Participants thereby tended to over-rely on AI, whereby explanations further increased false negative rates. The authors argue that explanations led to a reduced cognitive engagement with the task, which is in line with the findings of Bauer et al. (2023).	Scopus	No support AI	Why? Method: FA
16	Chen et al. (2022)	Predict whether furniture customisation will increase or decrease the price	The report employs three use-cases to explore which explanation method will support human task performance best. The relevant real-world task ('counterfactual') for this meta-analysis involved furniture arrangement, where participants had to decide whether increasing the length of a given furniture will increase its price. The human performance is thereby compared to simulated agents' performance. Participants and the agents received either no support or XAI-based support in the form of SHAP, LIME, or GAM explanations, which highlighted the contribution of the furniture size to a price increase or decrease. Results demonstrated that LIME explanations could significantly increase task performance when compared to no support, while SHAP and GAM explanations could not. The authors assume that this is because participants were more successful in learning how to use LIME explanations.	Scopus/Web of Science	No support	Why? Method: FA
31	Chiaburu et al. (2024)	Predict species of wild bees	Within the study of this report, participants had to annotated images of bee species using a visual DSS. All participants received no support in the first task and in the second varying support ranging from solely AI-based support in the form of predictions (with and without confidence ratings) or support with accompanying explanations. The explanations provided concept-based as well as EB explanations. In this meta-analysis we only considered XAI-based support in the form of EB explanations and predictions as AI-based support. Explanations improved participants' classification accuracy and reduced uncertainty, especially for challenging images. However, participants showed over-reliance, as they tended to blindly trust the AI's output when EB explanations are provided.	Forward Search	AI	What-else? Method: EB
20 21	Nguyen et al. (2021)	Classify objects and animals	In this report, participants classified images from two datasets (i.e. two studies; ImageNet, Stanford Dogs) and completed a training to familiarise with the task and a validation/test phase to test the impact of varying decision support provided. The support conditions included AI predictions and XAI in the form of FA and nearest-neighbour examples (EB explanations). The report found that XAI-based support did not consistently improve performance; in some cases, it reduced accuracy compared to AI-only predictions. More specifically, the results show that EB explanations were more effective than FA explanations.	Scopus/Web of Science	AI	Why?/What-else? Methods: FA/EB

(*Continued*)

(Continued).

Study ID(s)	Report	Task	Description	Source/Origin	Control conditions	Explanation type(s)
17 18	Gombolay et al. (2024)	Predict paediatric and neurological disorders	In two studies, participants performed neurology-related diagnostic tasks. The first study involved participants from the general population, whereas the second focused on a medical population. Participants received either predictions without explanations or with various explanation methods, including CF, EB (described as 'case-based' explanations) and FI explanations. Results indicated that the explanations did not consistently enhance decision accuracy. Higher experience and perceived explainability thereby seems to negatively impact decision performance. The authors conclude that there is no one-size-fits all approach, as the XAI support conditions are not uniformly beneficial.	Scopus	AI	Why?/What-else?/How-to? Methods: FA/EB/CF
28	Goyal et al. (2024)	Predict whether applicants will complete the loan on time	In this report, the authors explore how explanations influence fairness in decision-making when AI models contain bias. They distinguish between the impact of protected (e.g. sensitive attributes such as gender) and proxy features (e.g. not sensitive but highly correlated with protected features such as the university attendance). Participants had the the task to make loan approval decisions based on AI model predictions. The participants received AI predictions, either alone or paired with explanations that highlighted influential factors, such as employment status. While these explanations generally enhanced task performance, they also led participants to align more closely with the model's biases. For practical reasons, we combined groups within each proxy condition (with and without explanations) for the analysis.	Forward Search	AI	Why? Method: FA
33	Humer et al. (2024)	Decide whether a mushroom is edible and suitable for taking home	The study of this report investigates AI and XAI-based decision support in a mushroom hunting task. Participants identified poisonous and edible mushrooms using an (X)AI-assisted decision-making system. We calculated decision performance from the mean of the edible assessment and take-home decision, as both are task performance related. Participants received AI predictions either alone or paired with explanations, including FI explanations (E1/E2), and nearest-neighbour examples (i.e. EB explanations; E3). Explanations significantly improved task performance, while EB explanations proved to be the most effective overall. Additionally, all XAI conditions appeared effective in fostering trust in the AI when it provided correct answers.	Forward Search	AI	Why?/What-else? Methods: FI/EB
9	Jakubik et al. (2022)	Decide whether a borrower will make a repayment	The report focuses on decision performance and over-reliance in the context of peer-to-peer lending. Participants receive AI support in three forms: recommendations indicating whether repayment will occur, CF outcomes (referred to as 'AI with predicted outcomes'), or feature attribution explanations describing a borrower's loan application characteristics. The report yields no significant differences between the decision support conditions. The results show that participants tend to rely on AI predictions when participants receive CF, especially when these predictions are incorrect.	Forward Search	AI	Why?/How-to? Methods: FI/CF

(*Continued*)

(Continued).

Study ID(s)	Report	Task	Description	Source/Origin	Control conditions	Explanation type(s)
7	Leichtmann, Hinterreiter, et al. (2023)	Decide whether a mushroom is edible and suitable for taking home	Similar to Humer et al. (2024), participants receive an app with which they can decide whether a mushroom is poisonous and can be taken home. However, in this study, participant assess the mushrooms at a public art festival. The AI-based support provides a confidence rating, while the FA explanations use pixel attributions to highlight which areas of the image are relevant for a corresponding assessment. The results show that participants that received explanations outperform sole AI-based support. In addition, the authors find no effect of explanation on visitors' trust and acceptance. *Please note that, although the case is similar to that of Humer et al. (2024), the employed sample, decision support conditions, and study setting differ in the respective reports.*	Scopus	AI	Why? Method: FA
2	Leichtmann, Humer, et al. (2023)	Decide whether a mushroom is edible and suitable for taking home	The study of this report revolves around a task in mushroom hunting. In this study, the authors explore how AI and explainable AI (XAI) decision support impact performance in an online experiment. The findings reveal that explanations generally improved task performance and helped participants better calibrate their trust. Interestingly, prior educational interventions and domain-specific knowledge did not significantly influence task performance. *Please also note here that, although the case is similar to that of Humer et al. (2024) and Leichtmann, Hinterreiter, et al. (2023), the employed sample, decision support conditions, and study setting differ in the respective reports.*	Scopus/Web of Science	AI	Why? Method: FA
4	Liu et al. (2021)	Predicting if defendants violate their terms (ICSPR) and reoffend (COMPAS)/Predict the profession of persons (BIOS)	The report conducted two studies using in and out-of-distribution data. Following Schemmer et al. (2022), we focus on the in-of-distribution data setting to reduce between-study heterogeneity. The in-of-distribution study investigates the difference between static and interactive FA explanations in three simple prediction tasks. Static explanations simply show feature attributions whereas interactive explanations allow users to explore different scenarios, while the interface displays changes in the predictions and feature attributions. The results show that FA improve participants performance in the prediction of profession. In tasks that revolve around the violation of terms and the recidivism of defender, the results of the paper yield no significant difference between human performance and no support as well as XAI-based decision support.	Forward Search	No support	Why? Method: FA
24 25	Ma et al. (2024)	Predict whether defendants will recidivate in two years/Determine if a person has a good/bad credit rating	This report introduces 'OpenHEXAI', a framework for the human-centred evaluation of XAI methods. Participants evaluated decision tasks in two studies based on two datasets (German credit/RCDV) comparing the effect of no support, AI-only predictions, and XAI-based post hoc explanations on task performance. In summary, the study benchmarked four FA-based explanation methods (LIME, SHAP, IG, SG). Findings revealed that while XAI explanations enhanced trust and perceived fairness, their impact on decision accuracy varied depending on the specific explanation method.	Forward Search	No support AI	Why? Method: FA

(*Continued*)

(Continued).

Study ID(s)	Report	Task	Description	Source/Origin	Control conditions	Explanation type(s)
26	Maehigashi et al. (2024)	Identify rotten fruits	This report explores how the level of explanation impacts trust and task performance in visual tasks. In the study analysed, participants identified rotten strawberries among fresh ones. The experiment included both low-complexity and high-complexity tasks, evaluated within the same conditions. Task performance was measured as the average accuracy across both task types. Participants received different levels of support: no support, AI predictions only, explanations with heatmaps (FA explanations), or explanations supplemented by AI reliability and goal information. Results showed that AI predictions alone improved trust and performance in low-complexity tasks, while XAI with additional AI reliability and goal details further enhanced trust and performance in high-complexity tasks.	Forward Search	AI	Why? Method: FA
32	Wang and Ding (2024)	Predict sales of a product	This report explores trust and decision performance in human-XAI collaborations. Participants predicted whether the sales of a product will be higher than 10,000 pieces under AI-only predictions or support through FA-based explanations (either with rational or random explanations). To reduce between-study heterogeneity, we only considered rational explanations. The study found that XAI explanations improved decision accuracy and behavioural trust (i.e. consistency with the AI's decision) but did not enhance self-reported trust.	Forward Search	AI	Why? Method: FA
27	Papenmeier et al. (2022)	Identify offensive tweets	The report investigates the relationship between trust, ML model accuracy and explanations in AI. To this end, participants judged offensive tweets with support from an (X)AI system with varying accuracy. Three conditions were tested: no support, AI predictions with random or faithful explanations, where we only considered faithful explanations and a ML model accuracy of high and medium to reduce variability among studies. Results showed that explanation did not improve participants' task performance. Additionally, the findings show that the effect of XAI on user trust varies based on the accuracy of the AI model and the trust measure (i.e. self-reported trust vs. behavioural trust).	Backward Search	AI	Why? Method: FI
29	Reingold et al. (2024)	Identify deceptive hotel reviews	In this report, participants identified deceptive hotel reviews with (X)AI support under four conditions. AI support (C0; i.e. single-model predictions with explanations), and three forms of FA-based explanations: C1 highlights pos. FA for the prediction; C2 additionally included a both neg. and pos. FAs for the AI output; C3 showed supporting and opposing FA for a respective predicted class. The results of the paper yield no significant differences for task performance between the study conditions. However, the results demonstrated that dissenting explanations (C2) significantly reduced overreliance on AI predictions compared to the condition C1.	Forward Search	AI	Why? Method: FA
15	Silva et al. (2022)	Various tasks	The report explores the objective and subjective impact of human-XAI interactions. Participants must answer questions on various topics (e.g. decide on whether a pet requires feeding), relying on support from an AI-based support that offers predictions (i.e. no explanations) and various XAI-based support conditions. For our meta-analysis, we consider the FI and CF explanation conditions. The authors measure the objective impact through completion time, accuracy, and compliance with the AI's decision and subjective impact through trust, perceived social competence, and self-rated explainability. The findings of the study show that FI explanation indeed increased task performance, whereas CF explanations did not.	EBSCOHost/ Scopus/Web of Science	No support	Why?/How-to? Methods: FI/ CF

(*Continued*)

(Continued).

Study ID(s)	Report	Task	Description	Source/Origin	Control conditions	Explanation type(s)
19	Nguyen et al. (2023)	Predict toxicity in tweets	This report proposes an evaluation method to assess the fairness of explanations in AI and their impact on different user groups (Asian vs. non-Asian participants) in the context of hate speech. Participants had to decide whether tweets were hate speech directed at Asian people or not. They received AI predictions only or additional XAI-based explanations (FA and CF) in the form of highlighted words. Results showed that FA improved human task performance and reduced disparities in mental discomfort between different groups of moderators, while CF explanations were less effective and sometimes even caused mental discomfort.	Scopus	AI	Why?/How-to? Methods: FA/CF
10	Lai et al. (2020)	Identify if hotel reviews are fake	In this study, participants reviewed hotel text passages to determine whether the reviews were fake (i.e. bot-generated) or genuine. They either received no assistance or AI predictions accompanied by feature attribution (FA) explanations in three formats. The basic XAI condition highlighted specific text fragments suggesting a fake review. The second format built on this by incorporating guidelines from related research or the underlying AI model. The final format added performance-related statements about the AI. The authors found an overall but no significant increase in task performance for all XAI treatments.	Backward Search	No support	Why? Method: FA
3	van der Waa et al. (2021)	Estimate correct insulin dose	Within this report, participants are confronted with diabetic patients for which they have to choose the right insulin dose for a given situation. Participants receive either AI-support or XAI-based explanations in the form of EB and rule-based explanations. EB explanations show comparable situations from the past, where rule-based explanations display a certain outcome for given input data, which we consider as anchor explanations. Interestingly, the results show that participants tended to rely more on the AI's prediction when explanations are provided. The paper reports no significant difference between AI and XAI-based decision support.	Scopus/Web of Science	AI	What-else ?/How-to? Methods: EB CF/Anchor
30	Walter et al. (2023)	Classify the city of origin in Google Street View images	The report examines how XAI influences users' task performance. Participants were asked to identify the city of origin in Google Street View images across eight rounds. In each round, they were presented with one image to match to one of four possible cities. Participants were divided into three types of support conditions: no assistance, basic AI predictions, and XAI-based explanations accompanying predictions. The authors report that the AI-supported group outperformed the group without any decision support. Furthermore, the study demonstrates that providing explanations within AI-based decision support can substantially enhance task performance.	Forward Search	No support AI	Why? Method: FA
5 6	Wang and Yin (2021)	Predict recidivism of offenders and spruce-fir forest covering	This report examines the effectiveness of different XAI methods in human decision-making. Within two studies, the authors evaluate 4 XAI methods of three types. More specifically, participants engaged in recidivism and forest cover predictions under varying levels of domain expertise. Support conditions included AI predictions, and XAI-based explanations in the form of FA, FI, EB, and CF explanations. The results reveal that FI explanations were most effective for both forest cover and recidivism predictions.	Scopus	AI	Why?/What-else?/How-to? Methods: FA/FI/EB/CF

(*Continued*)

(Continued).

Study ID(s)	Report	Task	Description	Source/Origin	Control conditions	Explanation type(s)
8	Zhang et al. (2020)	Estimate if a person's income is smaller or larger than 50k dollars	The study of this report comprises three decision support conditions from two studies. In the first study, one group of participants receives basic AI support that predicts whether a subjects income exceeds 50k dollar, whereas the other group received additional AI-based confidence ratings. In the second study, participants were provided with FA explanations (SHAP) that show how the characteristics of a person contribute to the AI output. To reduce between-study heterogeneity, we only consider the sole AI decision support from the first study as the control group. The experiment shows that neither the XAI-based support nor the confidence scores improved task performance compared to AI-only support. However, confidence scores helped to calibrate trust in AI outputs.	Backward Search	AI	Why? Method: FA