Effectiveness of AI-based decision support systems in work environment: a systematic literature review

Katharina Buschmeyer*

Faculty of Business,
Augsburg Technical University of Applied Science,
Augsburg, 86163, Germany
Email: katharina.buschmeyer@tha.de
*Corresponding author

Julie Zenner

Faculty of Liberal Arts and Science,
Augsburg Technical University of Applied Science,
Augsburg, 86163, Germany
Email: zenner.ju@gmail.com

Sarah Hatfield

Faculty of Business,
Augsburg Technical University of Applied Science,
Augsburg, 86163, Germany
Email: sarah.hatfield@tha.de

Abstract: Artificial intelligence (AI) is being increasingly used in high-stakes working areas to augment experts in challenging decision-making situations. The AI support is intended to reduce the cognitive load on experts, which should ideally be reflected both in a greater sense of well-being when working on demanding tasks and in joint performance exceeding that of both the humans and AI alone. However, the extent and conditions of achievement (such as the AI accuracy and explainability) of these intended effects have not been systematically investigated. Therefore, we identified and reviewed 44 articles published since 2018 that have investigated the effects of AI-based decision support systems on experts in controlled experimental settings. The results suggest that, for optimal human-AI performance, which surpasses the performance of either alone, both must operate at similar and high levels. However, the effect on the psychological load remains unclear owing to limited research.

Keywords: artificial intelligence; decision making; AI-based decision support systems; cognitive relief; task performance.

Reference to this paper should be made as follows: Buschmeyer, K., Zenner, J. and Hatfield, S. (2024) 'Effectiveness of AI-based decision support systems in work environment: a systematic literature review', *Int. J. Human Factors and Ergonomics*, Vol. 11, No. 5, pp.1–54.

Biographical notes: Katharina Buschmeyer is a junior researcher in the field of Occupational Psychology at the Augsburg Technical University of Applied Sciences and a PhD student at the Ruhr University Bochum in Germany. Her research focuses on the use of augmented intelligence systems in the professional context to improve working conditions and the experience and behaviour of employees. She holds a Master's in Business Psychology in 2019.

Julie Zenner has a Doctoral in the field of Pedagogical Psychology from the University of Siegen in Germany in 2020. In her work, she focuses on the interaction between relationship experiences and career aspirations. Her current research interest is in the field of human-machine interaction at the Augsburg Technical University of Applied Sciences.

Sarah Hatfield is a Professor of Change Management and Human Resources at the Augsburg Technical University of Applied Sciences. She is the Founder and the Head of the university's BSc in Business Psychology and holds a diploma in work, organisation and business psychology. Her research focuses on human-AI interaction, learning in virtual reality settings, and gamified learning.

1 Introduction

In high-stakes working areas, such as finance, healthcare and law, artificial intelligence (AI) applications are being increasingly used to assist professionals in making demanding decisions (Lai et al., 2023; Zhou et al., 2023). For this, AI systems process and analyse all available (unstructured) information and data for a specific decision situation – a task that usually exceeds human information processing capabilities (Marois and Ivanoff, 2005) – and provide the core results to humans in the form of predictions or recommendations (Janiesch et al., 2021; Jarrahi, 2018; Murphy, 2012). Users of such AI-based decision support systems (AI-DSS) can decide whether to follow the system's advice. The human decision-making authority is essential for legal, ethical and safety reasons in areas, where the consequences of decisions can be devastating (Lai et al., 2023). This is because although AI-DSS – which are mostly based on machine learning (ML) models (Zhang et al., 2020; Chen et al., 2023) – can have impressively high predictive performance, the correctness of their advice cannot be guaranteed owing to their probabilistic character (Zhang et al., 2020). In other words, a residual uncertainty of their erroneousness always prevails. Furthermore, ML models are only as accurate as the historical data used to train them, and this data may contain, for example, input errors and biases (Vasconcelos et al., 2018; Zhang et al., 2020). Thus, experts should evaluate the results of AI-DSS based on critical thinking, intuition, domain knowledge, and experience and maintain control over the decision-making process and associated actions (Hellebrandt et al., 2021; Spector and Ma, 2019; Wilkens, 2020).

Ideally, this human-AI joint decision-making performance should exceed the individual performance of both, the human and AI system alone (Zhang et al., 2020; Bansal et al., 2021; Levy et al., 2021). However, according to Buçinca et al. (2021) and Liu et al. (2021), this aim is only partially achieved; both the groups referred to many experimental studies in which the human-AI joint decision-making performance significantly outperformed individual human performance, and only in rare cases exceeded that of AI alone (see, e.g., Bansal et al., 2021; Buçinca et al., 2020, 2021; Carton et al., 2020; Green and Chen, 2019a, 2019b; Lai et al., 2020; Lai and Tan, 2019; Lin et al., 2020; Poursabzi-Sangdeh et al., 2021; Wang and Yin, 2021; Zhang et al., 2020). Thus, although humans are influenced by AI-DSS, they occasionally face difficulty establishing appropriate trust in AI-based systems and mistakenly reject correct AI advice or follow incorrect AI advice (Liu, 2021). Due to Vasconcelos et al. (2023), the latter type of error, known as 'overtrust', has mainly been observed in empirical studies to date. Thus, the efficiency of human-AI cooperation depends strongly on AI-DSS accuracy.

However, the applicability of these findings to the use of AI-DSS in high-stakes work contexts is unclear. This is because almost all studies reported by Buçinca et al. (2021) and Liu et al. (2021) investigated the effects of AI-DSS in non-professional contexts or used laypersons as interaction partners of AI-DSS. In professional contexts, however, as already discussed, the experts are the intended users of AI-DSS and should have the necessary domain knowledge and experience to critically scrutinise the AI advice. According to current research results, the domain-specific knowledge, in human-AI interaction, equips humans with ability to recognise and reject incorrect AI-DSS advice (Gaube et al., 2021; Bayer et al., 2022; Dikmen and Burns, 2022). This is probably due to experts having a better mental model of the decision situation than laypersons, and therefore, being able to detect errors easily. Second, humans with a high level of expertise also have higher professional identification (Beijaard et al., 2000) and more confidence in their judgements than those with less expertise (Gaube et al., 2023), which presumably causes them to question and reject the advice of AI systems more critically. Therefore, we aimed to examine the findings of current research that specifically analyses the influence of AI-DSS in professional decision-making situations with experts as users based on the following questions: Do experts show appropriate trust in AI-DSS so that they outperform themselves and the AI system through its assistance? Or do experts also tend to overly rely on AI, which leads to human-AI performance depending strongly on AI system accuracy? Alternatively, is there evidence of 'undertrust', resulting in low dependence on AI performance or advice?

In addition to the central influence of the accuracy of AI-DSS on user experience and behaviour, many researchers are currently discussing the importance of the explainability of AI-DSS (von de Merwe et al., 2022; Lai et al., 2023). They assume that the black-box nature and associated lack of transparency of ML-based systems make it difficult for users to know when they should and should not follow AI advice (Jussupow et al., 2021). Therefore, they attempt to use AI methods, referred to in this context as explainable AI (XAI) methods, to explain the functioning of ML-based applications and investigate user reactions to these additional explanations (Arrieta et al., 2020). Previous studies with laypersons as interaction partners of AI-DSS observed varying effects of additional explanations. First, no effect on the human-AI interaction (see, e.g., Weerts et al., 2019; Zhang et al., 2020), second, a more calibrated trust in the systems, reflecting improved human-AI performance (see, e.g., Mercado et al., 2016; Stowers et al., 2020), and third, a worse human-AI performance (see, e.g., Bansal et al., 2021), as users interpreted the additional explanations as a general sign of competence and their presence alone increased trust in the AI system (Buçinca et al., 2021).

At this point, the expert reactions to the additional explanations are unclear. However, they can presumably better judge the plausibility of system explanations by comparing them with the familiar expert rules, and thus, demonstrate a higher level of appropriate trust. The rules developed by the AI system, which become visible through the system explanations, do not necessarily match those of experts, regardless of whether they are correct. This can increase the mistrust of the experts in the AI system.

Therefore, this study was aimed at investigating the effects of AI-DSS and its accuracy and explainability in professional decision-making situations, specifically on experts. We are interested in the effects on:

1. the behaviour of experts in the form of changes in performance
2. the psychological load experienced by them in decision-making situations, e.g., their mental effort.

This is because, in our modern working world, there is an increasing emphasis on promoting not only performance but also considering the well-being of employees (Cai et al., 2019; Finck et al., 2022; Langer et al., 2021; Singh et al., 2022). To achieve this goal, we aim to systematically collect data on current experiments with experts as AI-DSS interaction partners in professional decision situations and evaluate their summarised results in relation to specific research questions (RQ):

RQ1 How does the provision of AI-DSS in work-related decision situations influence the:

a performance behaviour

b psychological load experience of experts?

RQ1a Does human-AI collaboration improve the performance of experts in work-related decision situations compared to firstly their individual performance without AI-DSS and secondly the individual performance of the AI-DSS without expert validation?

RQ1b How does human-AI collaboration improve psychological load experienced by experts in work-related decision situations compared to their psychological load experience without AI-DSS?

RQ2 How do individual characteristics of AI-DSS, especially its accuracy and explainability, influence the psychological load experienced by and performance behaviour of experts in work-related decision situations?

2 Methodology

A systematic literature review was conducted, wherein the results of existing studies on the impact of AI-DSS in work-related decision situations on psychological load experienced by and performance behaviour of users were summarised. This review adheres to the guidelines of the preferred reporting items for systematic reviews and meta-analyses (PRISMA) statement guidelines (Page et al., 2021). Following the PRISMA flowchart, in this section, the methodological approach is described in three steps:

1. identifying relevant studies
2. selecting studies

3 analysing the included studies and synthesising the findings.

The first two steps are described in this section, and the third step is discussed in the results section.

2.1 Identifying relevant studies

In this study, we identified and extracted scientific journal articles addressing the relationship between the provision of AI-DSS in work-related decision scenarios and psychological load experienced by experts and associated performance behaviour. For this, we used five major electronic databases: Scopus, Web of Science, ACM Digital Library, IEEE Xplore, and EBSCOhost (PsycINFO, PsycArticles, and PSYNDEX). To achieve this, we first identified a set of keywords related to the RQs (Table 1).

Table 1 Search string

Composite independent variable		*Context*	*Dependent variable*
Part 1	*Part 1*		
‘artificial intelligence’ OR ‘augmented intelligence’ OR ‘intelligence augmentation’ OR ‘AI’ OR ‘data driven’ OR ‘machine learning’	(decision NEAR/2 aid) OR (decision NEAR/2 assistan*) OR (decision NEAR/2 agent*) OR (decision NEAR/2 support*) OR (decision NEAR/2 system*)	work* OR job* OR employe* OR profession* OR occupation*	‘user experience’ OR behavio* OR *load OR stress OR mental OR psych* OR cognitiv OR perform* OR satis* OR confiden*

Notes: Example strings used in the Web of Science. In other databases, the operators were adapted as necessary, such as in Scopus: ‘W/2’ was used instead of ‘NEAR/2’. In all database searches, the four categories were linked with the Boolean operator ‘AND’.

A literature review completed in April 2024 yielded 10,917 relevant articles after filtering out papers not published in academic journals or proceedings, those not in English, and those published before 2018 (Table 2). The decision to include only recent studies in the review is based on the recent advances made in AI (Lai et al., 2023; Levy et al., 2021; Nicodeme, 2020). These developments have presumably led to current expectations that AI systems are significantly more powerful than non-AI-based applications (Almarashda et al., 2022), and can provide significant relief (Hornung and Smolnik, 2022). These expectations and attitudes influence human experience and behaviour in human-AI interaction (Ajzen et al., 2018; Liu et al., 2023); therefore, current AI users probably experience themselves differently when interacting with AI and behave differently than they did years ago. The decision to focus specifically on studies from 2018 onwards is based on a recent review by Lai et al. (2023). According to this review, research on human decision-making in the context of AI has increased significantly since 2018, with the advancements in AI technologies. Following the database search, we used the snowball sampling system (Wohlin, 2014) to explore suitable articles. This search yielded 57 articles. A total of 10,974 studies were identified, including 2,342 duplicate studies. Ultimately, 8,632 studies were included.

2.2 Study selection

The relevant studies were selected in two screening steps using the inclusion and exclusion criteria listed in Table 2. First, titles, abstracts, and keywords were checked, and unsuitable studies were eliminated. After thorough reading, the remaining studies were classified into 'include', 'exclude', and 'maybe' categories. Two independent reviewers conducted both the steps. The free Rayyan platform for systematic literature reviews (https://www.rayyan.ai/) was used for this process. Thereafter, the reviewers discussed the studies categorised under 'maybe' and 'include' by only one reviewer. In cases of persistent disagreements, a third reviewer was consulted. Finally, a consensus was reached in all cases.

Table 2 Inclusion and exclusion criteria for the review

		Inclusion criteria		*Exclusion criteria*	
Soft factors	Population and problem	Experts who are tasked with making work-related decisions		Laypersons with no expertise in the concerned task, which is often reflected in a lack of qualifications; people who do not have to make decisions or those whose decisions are not professional; people who make decisions in groups and not alone	
	Intervention	RQ1	Provision of an AI-DSS for decision making.	RQ1	Provision of a fully automated AI-based system or a conventional DSS that are not based on ML methods.
		RQ2	Provision of an AI-DSS for decision-making, focussing on the system design criteria of accuracy and explainability or to their extent.	RQ2	Same criteria as for RQ1; no focus on the two design criteria or their design is not considered from a generally valid perspective, but from a technical perspective, for example by comparing results of different XAI explanation methods, such as LIME and SHAP.
	Control	RQ1	A control group that is not provided with an AI-DSS (e.g., rule-based DSS or no DSS).	RQ1	No control group included; control group that does not relate to the system but to the experience level of the subjects

Notes: AI = artificial intelligence; DSS = decision support system; ML = machine learning; LIME = local interpretable model-agnostic explanations; SHAP = Shapley additive explanations.

Table 2 Inclusion and exclusion criteria for the review (continued)

		Inclusion criteria		*Exclusion criteria*	
Soft factors	Control	RQ2	A control group that is provided with an AI-DSS with a different design of the system characteristics than the intervention group.	RQ2	No control group included; control group that does not relate to the system but to the experience level of the subjects
	Outcome	Psychological load experienced by subjects (e.g., mental effort) and associated performance behaviour, which is measurable, and thus, comparable		Subject attitudes towards an AI-DSS, e.g., whether they rate it as trustworthy; qualitative individual statements about the experience of psychological load and related behaviours of professionals during tasks, which are difficult to compare	
Hard factors	Year of publication	Published in 2018 or later		Published before 2018	
	Language	English		Other languages, for, e.g., Spanish, Chinese, Korean, etc.	
	Publication type	Journals; conference papers; proceedings		Book chapters; magazine articles; reports; theses; dissertation	

Notes: AI = artificial intelligence; DSS = decision support system; ML = machine learning; LIME = local interpretable model-agnostic explanations; SHAP = Shapley additive explanations.

Figure 1 Flow diagram of the screening process

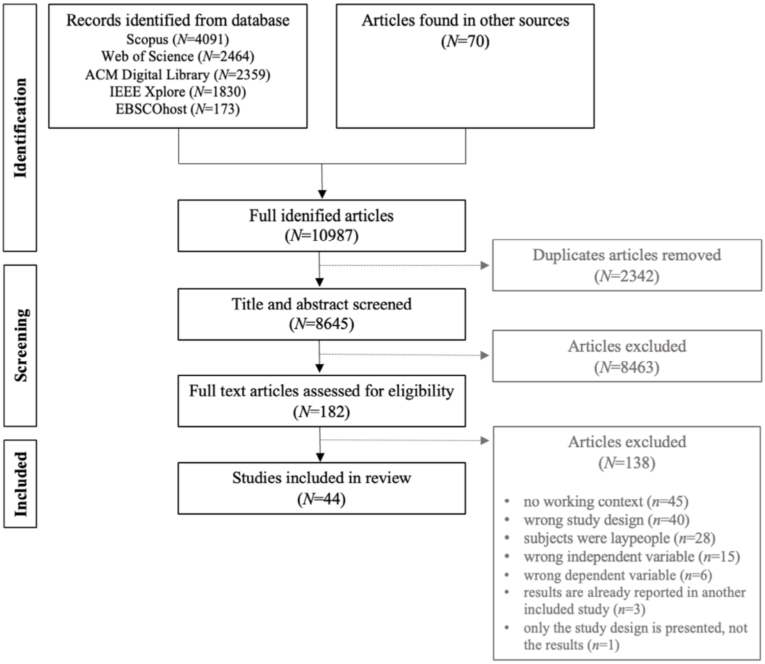
Identification
Records identified from database
Scopus (N=4091)
Web of Science (N=2464)
ACM Digital Library (N=2359)
IEEE Xplore (N=1830)
EBSCOhost (N=173)
Articles found in other sources
(N=70)
Full idenified articles
(N=10987)
Duplicates articles removed
(N=2342)
Screening
Title and abstract screened
(N=8645)
Articles excluded
(N=8463)
Full text articles assessed for eligibility
(N=182)
Included
Studies included in review
(N=44)
Articles excluded
(N=138)
• no working context (n=45)
• wrong study design (n=40)
• subjects were laypeople (n=28)
• wrong independent variable (n=15)
• wrong dependent variable (n=6)
• results are already reported in another included study (n=3)
• only the study design is presented, not the results (n=1)

Table 3 Descriptive analysis of the included studies

						Investigated variables					*Experimental research design*					
	Field of investigation					*Independent variables*					*Location*			*Methods*		
Authors					*Sample*	*AI-DSS in total*	*System characteristics*		*Dependent variables*	*Real vs. simulated investigated AI-DSS*						
	Medicine	*Aviation*	*Recruiting*	*Financial crimes*			*Accuracy*	*Explainability*			*Labour*	*Online*	*Field*	*Within-subject design*	*Between-subject design*	*Randomisation*
Didimo et al. (2018)				x	N = 32 workers at the IRV's fiscal audit office	x			Performance	Real	x				x	Yes
Rodriguez-Ruiz et al. (2018)	x				N = 7 physicians	x			Performance	Real	x			x		No
Cai et al. (2019)	x				N = 12 physicians	x			Workload	Real	x			x		Yes
Liu et al. (2019)	x				N = 2 physicians	x			Performance	Real	NR	NR	NR	x		Yes
Bai et al. (2020)	x				N = 6 physicians	x			Performance	Real	x			x		NR
Dorr et al. (2020)	x				N = 54 physicians	x			Performance	Real		x		x		Yes
Kiani et al. (2020)	x				N = 11 physicians	x	x		Performance	Real	x			x		Yes
Kim et al. (2020)	x				N = 3 physicians	x			Performance	Real	x			x		Yes
Kozuka et al. (2020)	x				N = 2 physicians	x			Performance	Real	x			x		Yes
Lee et al. (2020)	x				N = 5 therapists	x			Performance; workload	Real	x			x		Yes
Rajpurka et al. (2020)	x				N = 13 physicians	x			Performance	Real	NR	NR	NR	x		Yes

Notes: NR = not reported; IRV = Italian Revenue Agency.

Table 3 Descriptive analysis of the included studies (continued)

						Investigated variables					*Experimental research design*					
	Field of investigation					*Independent variables*					*Location*			*Methods*		
							System characteristics									
Authors	*Medicine*	*Aviation*	*Recruiting*	*Financial crimes*	*Sample*	*AI-DSS in total*	*Accuracy*	*Explainability*	*Dependent variables*	*Real vs. simulated investigated AI-DSS*	*Labour*	*Online*	*Field*	*Within-subject design*	*Between-subject design*	*Randomisation*
Gaube et al. (2021)	x				N = 265 novice physicians		x		Performance; confidence in the decision made	Simulated		x		x		Yes
Jacobs et al. (2021)	x				N = 220 physicians	x	x	x	Performance; confidence in the decision made	Simulated		x		x		Yes
Jussupow et al. (2021)	x				N = 47 physicians	x	x		Performance	Simulated	x			x		Yes
Kavya et al. (2021)	x				N = 4 physicians	x			Performance	Real	x			x		NR
Koo et al. (2021)	x				N = 4 physicians	x			Performance	Real	x			x		Yes
Martini et al. (2021)	x				N = 6 physicians	x			Performance	Real	x			x		Yes
Nam et al. (2021)	x				N = 6 physicians	x			Performance	Real	x			x		Yes
Popescu et al. (2021)	x				N = 7 physicians	x			Performance	Real			x	x		No
Rudie et al. (2021)	x				N = 7 physicians	x			Performance	Real	x			x		Yes
Singh et al. (2021)	x				N = 2 physicians	x			Performance	Real	x			x		Yes

Notes: NR = not reported; IRV = Italian Revenue Agency.

Table 3 Descriptive analysis of the included studies (continued)

						Investigated variables					*Experimental research design*					
	Field of investigation					*Independent variables*					*Location*			*Methods*		
Authors					*Sample*	*AI-DSS in total*	*System characteristics*		*Dependent variables*	*Real vs. simulated investigated AI-DSS*						
	Medicine	*Aviation*	*Recruiting*	*Financial crimes*			*Accuracy*	*Explainability*			*Labour*	*Online*	*Field*	*Within-subject design*	*Between-subject design*	*Randomisation*
Sung et al. (2021)	x				N = 6 physicians	x			Performance	Real	x			x		Yes
Yang et al. (2021)	x				N = 3 physicians	x			Performance	Real	x			x		NR
Yao et al. (2021)	x				N = 358 physicians	x			Performance	Real			x		x	Yes
Zhou et al. (2021)	x				N = 10 PhD students in industrial engineering, specialising in human factors and ergonomics	x			Performance; confidence in the decision made	Real		x		x		Yes
Calisto et al. (2022)	x				N = 45 physicians	x			Performance	Real	x			x		No
Duchevet et al. (2022)		x			N = 7 pilots	x			Performance; workload; safety feeling of the pilot	Real	x			x		Yes
Finck et al. (2022)	x				N = 4 physicians	x			Performance; confidence in the decision made	Real	x			x		No

Notes: NR = not reported; IRV = Italian Revenue Agency.

Table 3 Descriptive analysis of the included studies (continued)

						Investigated variables					*Experimental research design*					
	Field of investigation					*Independent variables*					*Location*			*Methods*		
							System characteristics									
Authors	*Medicine*	*Aviation*	*Recruiting*	*Financial crimes*	*Sample*	*AI-DSS in total*	*Accuracy*	*Explainability*	*Dependent variables*	*Real vs. simulated investigated AI-DSS*	*Labour*	*Online*	*Field*	*Within-subject design*	*Between-subject design*	*Randomisation*
Henkel et al. (2022)	x				N = 10 physicians	x			Performance	Real			x	x		No
Hwang et al. (2022)	x				N = 9 polysomnographic technicians	x		x	Performance	Real	x			x		Yes
Kiefer et al. (2022)	x				N = 7 physicians	x			Performance	Real	x			x		NR
Lacroux and Martin-Lacroux (2022)			x		N = 416 experienced person in personnel selection	x	x		Performance	Simulated		x			x	Yes
Panigutti et al. (2022)	x				N = 28 healthcare providers (physicians, nurses, healthcare assistant, dietetic assistant practitioner, ambulance call dispatcher)			x	Confidence in the decision made x	Real		x		x		Yes
Roller et al. (2022)	x				N = 8 physicians	x			Performance	Real	x			x		No
Gaube et al. (2023)	x				N = 223 physicians			x	Performance; confidence in the decision made	Simulated		x		x		Yes

Notes: NR = not reported; IRV = Italian Revenue Agency.

Table 3 Descriptive analysis of the included studies (continued)

Authors	Field of investigation				Sample	Investigated variables				Real vs. simulated investigated AI-DSS	Experimental research design					
						Independent variables			Dependent variables		Location			Methods		
						AI-DSS in total	System characteristics									
	Medicine	Aviation	Recruiting	Financial crimes			Accuracy	Explainability			Labour	Online	Field	Within-subject design	Between-subject design	Randomisation
Kim et al. (2023)	x				N = 49 nurses	x			Performance	Real			x		x	No
Kindler et al. (2023)	x				N = 2 physicians	x			Performance	Real	x			x		No
Laursen et al. (2023)	x				N = 13 physicians	x			Performance	Real	x			x		No
Lee and Chew (2023)	x				N = 7 therapists	x		x	Performance; workload	Real		x		x		Yes
Pushparaj et al. (2023)		x			N = 12 air traffic controller			x	Performance; workload	Real	x			x		No
Shah et al. (2023)	x				N = 10 physicians	x			Performance; confidence in the decision made	Real	x			x		NR
Sivaraman et al. (2023)	x				N = 24 physicians			x	Confidence in the decision made	Real		x		x		No
Yoon et al. (2023)	x				N = 66 physicians	x		x	Performance	Real		x		x		No
Zhang et al. (2023)	x				N = 2 physicians	x			Performance	Real	NR	NR	NR		x	Yes

Notes: NR = not reported; IRV = Italian Revenue Agency.

The selection process is illustrated in Figure 1. Initially, 8,645 studies were screened based on title, abstract and keywords, with 8,463 excluded as irrelevant. Subsequently, the remaining 182 articles were reviewed in terms of full text, resulting in 138 studies being excluded from further analysis. Among these was a paper by Li et al. (2021), which lists 38 studies on the impact of AI-DSS on radiologists in the detection of thoracic pathologies. A total of 13 of these 38 studies (Bai et al., 2020; Dorr et al., 2020; Kim et al., 2020; Koo et al., 2021; Kozuka et al., 2020; Liu et al., 2019; Martini et al., 2021; Nam et al., 2021; Rajpurkar et al., 2020; Singh et al., 2021; Sung et al., 2021; Yang et al., 2021; Zhang et al., 2023) met our inclusion and exclusion criteria and were included in our analysis (see Figure 1, 'articles found in other sources'). Finally, 44 peer-reviewed journal articles were included.

3 Results

3.1 Descriptive analysis of the included studies

The majority (n = 40) of the 44 studies examined the use of AI-DSS in a medical work context; thus, the participants in the identified studies were mostly physicians (n = 36). In total, 2,034 professionals participated in the 44 experiments, with an average cohort size of M = 46 individuals (SD = 95). The large standard deviation indicates the considerable variation in sample size, with subjects ranging from N = 2 (Kindler et al., 2023; Kozuka et al., 2020; Liu et al., 2019; Singh et al., 2021; Zhang et al., 2023) to N = 416 (Lacroux and Martin-Lacroux, 2022; see Table 3).

A total of 39 studies examined the general effect of AI-DSS in work-related decision scenarios on their users and answered RQ1 (Table 3). They compared the experience and behaviour of subjects when performing a (simulated) work task with and without an AI-DSS. Under control conditions (without AI-based support), the subjects received no technical support in most cases; only in a few individual cases was some other form of technical support provided, e.g., by conventional software systems (Didimo et al., 2018; Lee et al., 2020). Most of the included studies used a within-subject design (n = 38), wherein the order of the experimental conditions (with vs. without AI support) was randomly assigned.

Twelve studies included in this review examined the effects of the individual characteristics of AI-DSS and answered the RQ2. Eleven of the nine studies used a within-subjects design wherein the participants were randomly exposed to all conditions (Table 3).

3.2 Results for RQ1: effects of the provision of AI-DSS in work-related decision situations on

3.2.1 Performance of experts compared to their individual performance without AI-DSS

Of the 39 studies that examined the general influence of AI-DSS usage in work-related decision-making scenarios, all but one (Cai et al., 2019) explored how AI affects the task performance of users. Most of these studies investigated this by examining whether

AI-DSS usage improves the accuracy of decisions made during task processing (n = 16) and/or decreases the time required to complete the tasks (n = 14; see Table 4).

Of the 16 studies that examined the impact of AI-DSS on the task accuracy of users, 12 reported a recognisable improvement (see Table 4). However, only seven studies indicated a significant difference (Bai et al., 2020; Didimo et al., 2018; Finck et al., 2022; Nam et al., 2021; Rajpurkar et al., 2020; Rudie et al., 2021; Zhou et al., 2021). The remaining five studies did not report significance values (Kavya et al., 2021; Laursen et al., 2023; Yang et al., 2021; Yoon et al., 2023; Zhang et al., 2023). This is probably owing to the often very small sample sizes, which ranged from N = 2 (Zhang et al., 2023) to N = 36 (Yoon et al., 2023). Therefore, to understand the influence of the AI support in work-related decision-making scenarios better, we calculated the effect sizes of the changes in task accuracy using Cohen's d for all studies (see Table 5) except one by Zhou et al. (2021), because no specific accuracy values was reported therein. According to Cohen (1988), a d-value between 0.2 and below 0.5 is considered a small effect, a d-value between 0.5 and below 0.8 is considered a medium effect, and a d-value above 0.8 is considered a large effect.

Of the remaining 15 studies, three showed a strong effect of AI-DSS on the task performance of users (see Table 5). The largest effect was identified by Laursen et al. (2023), with d = 4.45. In their experiment, physicians searched a patient record for haemorrhages within a given time, first without and then with AI support. The AI highlighted relevant text passages with high sensitivity (93.7%) and specificity (98.1%), and participants were informed of the system performance beforehand. However, as there was no washout period between conditions, practice effects cannot be ruled out. Didimo et al. (2018), who used a different study design from that of Laursen et al. (2023), with a between-subject approach, also demonstrated a strong effect of AI assistance. For example, tax authority employees improved their task performance in one of the two task sets from 63.09% to 98.83%, corresponding to d = 3.27. The authors did not mention whether the participants were informed about the AI performance beforehand, and no specific data on the system performance were provided, other than that indicating that it was a high-performing system. In the study by Yoon et al. (2023), for which we calculated the third strongest effect, the AI alone achieved a 96.3% task accuracy, which was 4.5% higher than the baseline performance of the N = 36 ophthalmologists. With AI support, their performance increased to 95.2%.

Four studies demonstrated a moderate performance improvement through the use of AI assistance (see Table 5). Notably, in three of these studies (Bai et al., 2020; Rajpurkar et al., 2020; Rudie et al., 2021), similar to the results of Yoon et al. (2021), the AI performance surpassed that of human experts. However, in the study of Rudie et al. (2021), this was only the case in a subset of radiology residents, where a significant performance increase from 30% to 55% was noted. Among senior radiologists, their own performance (69%) was higher than that of the AI (61%), and no significant effect was observed, with their task accuracy only improving slightly to 72%. As opposed to the other three studies, Zhang et al. (2023) did not report separate AI performance. However, it is important to note that this study involved only two participants and used a between-subject design, making it unclear whether the observed differences were owing to the system or individual competency variations.

For the study by Yang et al. (2021), we calculated an effect size of d = 0.33, indicating a small effect. The task performance was already high without the AI system for the three radiologists at 94.1%, and with AI support (91.4% individual AI performance), it increased slightly to 95.1%. It is worth noting that the within-subject design did not include a washout period.

In the remaining seven studies, three investigations (Jussupow et al., 2021; Kiani et al., 2020; Lacroux and Martin-Lacroux, 2022) demonstrated that when the standalone performance of the AI was lower than that of the human participants, the AI assistance according to Cohen's d had no significant impact on the average performance of experts. This was consistent with findings from the sample of experienced radiologists in Rudie et al. (2021). However, it is notable that, despite the lack of a statistically measurable effect, the task performance in two of these cases actually declined owing to AI support, which had an accuracy of 50% in both instances. For example, Lacroux and Martin-Lacroux (2022) examined the effects of an AI-DSS on personnel selection, which was intended to assist in selecting the most suitable candidate for a position. They found that expert performance declined with AI support, as the accuracy of their hiring decisions decreased from 64.2% to 56.1%. It is also interesting to note that, in the studies by Kavya et al. (2021) and Jacobs et al. (2021), the standalone performance of the AI significantly exceeded that of the human participants, yet in both cases, the human performance did not improve substantially with AI assistance. In the case of Jacobs et al. (2021), this could be attributed to the nature of the experimental task, where participants were asked to make medical treatment decisions regarding antidepressants – a field in which opinions on the correct course of treatment often vary significantly. Finally, the study by Finck et al. (2022) highlighted that the effect of AI support largely depends on the potential for improvement. In this case, human performance without AI was already at 96.6% and increased to 99.1% with AI assistance, resulting in an effect size of d = 0.18.

Of the 14 studies examining the impact of AI-DSS on the task processing time of users, 11 reported a noticeable improvement. However, only six studies indicated a statistically significant difference. The remaining five studies did not report significance values, likely owing to the often very small sample sizes, as mentioned above (see Table 4). Three studies found no increase in performance in terms of efficiency, i.e., the time required for completing tasks (Kiefer et al., 2022; Lee et al., 2020; Shah et al., 2023). Notably, these studies investigated AI-DSS with extensive functionalities and interfaces. We calculated the effect sizes of the changes with and without AI using Cohen's d to gain a deeper understanding of the impact of AI support on task processing time in work-related decision-making scenarios. This calculation was possible for nine studies (see Table 6), as five studies did not report the necessary mean values and standard deviations (Henkel et al., 2022; Kiefer et al., 2020; Lee et al., 2022; Liu et al., 2019; Zhang et al., 2023). Of the nine studies for which we calculated effect sizes, four showed a large effect according to Cohen (1988), two showed a medium effect, one showed a small effect, and two showed no effect (see Table 6). It is important to note that, with one exception (Calisto et al., 2022), all studies that used a within-subject design (see Table 3) included a washout period, such as four weeks (Finck et al., 2022), making practice effects for time reduction unlikely.

Table 4 Overview of the impact of AI-DSS use on the task performance of experts, listed by publication date

Study	*Detailed sample description*	*Experimental task*	*AI-DSS (in the experimental condition)*	*Decision quality in terms of*					*Decision efficiency in terms of*	
				Accuracy	*AUC*	*Sensitivity*	*Specificity*	*Other criteria*	*Average processing time*	*Other criteria*
Didimo et al. (2018)	N = 32 participants who work in the fiscal audit office of the Italian Revenue Agency (IRV)	Detection of tax evasion	System assists to process the task by, for example, visually defining classes of suspicious patterns, based both on topological properties and on node/ edge attributes.	↑**					↓**	
Rodriguez-Ruiz et al. (2018)	N = 7 radiologist	Medical diagnosis (breast cancer)	System displays a cancer probability score for a specific area previously selected by the radiologist on the mammogram.		↑*	↑ (n.s.)	↑*			
Liu et al. (2019)	N = 2 radiologists	Medical diagnosis (pulmonary nodules)	System marks identified nodules with a square bounding box and also indicates the identified nodule type and the model confidence in its prediction.		↑ (sig. NR)				↓ (sig. NR)	
Bai et al. (2020)	N = 6 radiologists	Medical diagnosis (COVID-19 pneumonia)	System identifies if a chest CT shows COVID-19 or non-COVID pneumonia slices (87% accuracy).	↑**		↑**	↑**			

Notes: AUC: area under the receiver operating characteristic curve; ↑ = increase; ↓ = decrease; → = stable; no arrow means no direction is reported; ** means $p \leq 0.01$; * means $p \leq 0.05$; n.s. means $p \geq 0.05$; sig. NR = not reported if the difference is significant.

Table 4 Overview of the impact of AI-DSS use on the task performance of experts, listed by publication date (continued)

Study	*Detailed sample description*	*Experimental task*	*AI-DSS (in the experimental condition)*	*Decision quality in terms of*					*Decision efficiency in terms of*	
				Accuracy	*AUC*	*Sensitivity*	*Specificity*	*Other criteria*	*Average processing time*	*Other criteria*
Dorr et al. (2020)	N = 54 physicians	Medical diagnosis (COVID-19 pneumonia)	System identifies patients with COVID-19 pneumonia, viral and bacterial pneumonia, or no pneumonia (0.83 AUC) based on chest CT of patients.			↑**	↓**			
Kiani et al. (2020)	Study 1: N = 11 pathologists	Medical diagnosis (liver cancer)	Systems helps distinguish between the two most common types of primary liver cancer by giving a probability for each diagnosis with a class activation map to help with interpretation (84.2% accuracy).	↑ (sig. NR)						
Kim et al. (2020)	N = 3 emergency department physicians	Medical diagnosis (detection of visible pneumonia on chest radiographs (CR)	System provided a probability score for the presence of the aforementioned thoracic diseases and created a heat map of the input CR to facilitate the localisation of the lesion (0.940 AUC).		↑**	↑*	↑**		↓ (sig. NR)	

Notes: AUC: area under the receiver operating characteristic curve; ↑ = increase; ↓ = decrease; → = stable; no arrow means no direction is reported; ** means $p \leq 0.01$; * means $p \leq 0.05$; n.s. means $p \geq 0.05$; sig. NR = not reported if the difference is significant.

Table 4 Overview of the impact of AI-DSS use on the task performance of experts, listed by publication date (continued)

Study	*Detailed sample description*	*Experimental task*	*AI-DSS (in the experimental condition)*	*Decision quality in terms of*					*Decision efficiency in terms of*	
				Accuracy	*AUC*	*Sensitivity*	*Specificity*	*Other criteria*	*Average processing time*	*Other criteria*
Kozuka et al. (2020)	N = 2 radiologists (1–4 years of experience)	Medical diagnosis (detection of pulmonary nodules)	System assists in the detection of pulmonary nodules in CT images and has several functions like displaying marks, density, major axis, and volume of the detected nodules.			↑**	↓ (sig. NR)		↓ (sig. NR)	
Lee et al. (2020)	N = 5 therapists	Rehabilitation assessment	System predicts the quality of motion of a patient based on the rehabilitation exercises of the patient based on video data and generate user-specific analysis that includes feature analysis, images of salient frames, and graphs of joint trajectories.						↑ (n.s.)	

Notes: AUC: area under the receiver operating characteristic curve; ↑ = increase; ↓ = decrease; → = stable; no arrow means no direction is reported; ** means $p \leq 0.01$; * means $p \leq 0.05$; n.s. means $p \geq 0.05$; sig. NR = not reported if the difference is significant.

Table 4 Overview of the impact of AI-DSS use on the task performance of experts, listed by publication date (continued)

Study	*Detailed sample description*	*Experimental task*	*AI-DSS (in the experimental condition)*	*Decision quality in terms of*					*Decision efficiency in terms of*	
				Accuracy	*AUC*	*Sensitivity*	*Specificity*	*Other criteria*	*Average processing time*	*Other criteria*
Rajpurkar et al. (2020)	N = 13 physicians with anywhere from 6 months to 25 years of experience diagnosing tuberculosis (TB) in patients with HIV in South Africa	Medical diagnosis (tuberculosis)	System estimates probability of patient having active pulmonary TB and displays the result in five categories from very unlikely to very likely. In addition, the auxiliary interface also includes an explanation of the prediction, highlighting the areas of the X-ray that are most likely to indicate TB according to the algorithm (79% accuracy).	↑*	↑ (sig. NR)	↑ (sig. NR)				
Jacobs et al. (2021)	N = 220 physicians which have >1 year experience prescribing antidepressant treatments	Medical treatment decisions (antidepressant)	System provides treatment advice for major depressive disorder (66.7% accuracy for top diagnosis).	↓ (n.s.)						

Notes: AUC: area under the receiver operating characteristic curve; ↑ = increase; ↓ = decrease; → = stable; no arrow means no direction is reported; ** means $p \leq 0.01$; * means $p \leq 0.05$; n.s. means $p \geq 0.05$; sig. NR = not reported if the difference is significant.

Table 4 Overview of the impact of AI-DSS use on the task performance of experts, listed by publication date (continued)

Study	*Detailed sample description*	*Experimental task*	*AI-DSS (in the experimental condition)*	*Decision quality in terms of*					*Decision efficiency in terms of*	
				Accuracy	*AUC*	*Sensitivity*	*Specificity*	*Other criteria*	*Average processing time*	*Other criteria*
Kavya et al. (2021)	N = 4 physicians	Medical diagnosis (allergy)	System assists in diagnosing of coexisting allergic disorders and provides reasoning behind the predictions using post-hoc XAI approaches (86.39% accuracy).	↑ (sig. NR)						
Koo et al. (2021)	N = 4 radiologists (N = 2 thoracic radiologists and N = 2 radiological residents)	Medical diagnosis (pulmonary nodules)	Software identified regions of interest and awarded anomaly scores of 0–100% (0.87 AUC).		↑**	↑ (sig. NR)	↑ (sig. NR)			
Martini et al. (2021)	N = 6 radiologists	Medical diagnosis (pulmonary nodules)	System assists with detection of pulmonary nodules.						↓**	
Nam et al. (2021)	N = 6 radiologists (N = 2 thoracic radiologists, N = 2 board-certified radiologists, and N = 2 residents)	Medial diagnosis (e.g., pneumonia, pulmonary oedema, active tuberculosis)	Systems assist with detection of 10 common abnormalities on chest radiographs (0.895–1.00 AUC).	↑*					↓*	

Notes: AUC: area under the receiver operating characteristic curve; ↑ = increase; ↓ = decrease; → = stable; no arrow means no direction is reported; ** means $p \leq 0.01$; * means $p \leq 0.05$; n.s. means $p \geq 0.05$; sig. NR = not reported if the difference is significant.

Table 4 Overview of the impact of AI-DSS use on the task performance of experts, listed by publication date (continued)

Study	*Detailed sample description*	*Experimental task*	*AI-DSS (in the experimental condition)*	*Decision quality in terms of*					*Decision efficiency in terms of*	
				Accuracy	*AUC*	*Sensitivity*	*Specificity*	*Other criteria*	*Average processing time*	*Other criteria*
Popescu et al. (2021)	N = 7 physicians (including family physicians and psychiatrists)	Medical treatment decisions (major depressive disorder)	System support treatment selection for by providing individualised probabilities of remission for specific treatment options.							Patient appointment length: (n.s.)
Rudie et al. (2021)	Study 1: N = 4 radiology residents	Medical diagnosis (on multimodal brain MRI)	System provides, e.g., diagnostic advice (61% accuracy).	Study 1: ↑**						
	Study 2: N = 3 neuroradiologists			Study 2: ↑ (n.s.)						
Singh et al. (2021)	N = 2 radiologists	Medical diagnosis (subsolid nodules)	System assists with detection of pulmonary nodules.		↑ (n.s.)					
Sung et al. (2021)	N = 6 (N = 2 thoracic radiologists, N = 2 board-certified radiologists, N = 1 radiology resident, and N = 1 non-radiology resident)	Medical diagnosis (detecting and localising major abnormal findings like nodules on chest radiographs)	System identifies nodules, consolidation, interstitial opacity, pleural effusion, and pneumothorax.		↑*	↑*	↑*		↓**	

Notes: AUC: area under the receiver operating characteristic curve; ↑ = increase; ↓ = decrease; → = stable; no arrow means no direction is reported; ** means p ≤ 0.01; * means p ≤ 0.05; n.s. means p ≥ 0.05; sig. NR = not reported if the difference is significant.

Table 4 Overview of the impact of AI-DSS use on the task performance of experts, listed by publication date (continued)

Study	*Detailed sample description*	*Experimental task*	*AI-DSS (in the experimental condition)*	*Decision quality in terms of*					*Decision efficiency in terms of*	
				Accuracy	*AUC*	*Sensitivity*	*Specificity*	*Other criteria*	*Average processing time*	*Other criteria*
Yang et al. (2021)	N = 3 radiologists	Medical diagnosis (distinguishing COVID-19 infected pneumonia patients from fnon-COVID-19 infected patients on CT scans)	System diagnose COVID-19 using chest CT images of different types of pulmonary diseases, including tuberculosis, common pneumonia, non-COVID19 viral pneumonia, and COVID-19 pneumonia (0.903 AUC; 91.8% sensitivity and 91.4% accuracy).	↑ (sig. NR)		↑ (sig. NR)				
Yao et al. (2021)	N = 358 physicians	Medical diagnosis (low ejection fraction)	System identifies patients at a high likelihood of low EF based on a standard 12-lead electrocardiogram (system C-statistic of 0.92).					Detection rate of the disease ↑*		
Zhou et al. (2021)	N = 10 individuals who have or are currently pursuing PhD degrees in industrial engineering	Assess risks in lifting tasks faced by workers	Video-based system with prediction and explanation modules for assessing lifting risks (84.6% accuracy for top prediction).	↑*						

Notes: AUC: area under the receiver operating characteristic curve; ↑ = increase; ↓ = decrease; → = stable; no arrow means no direction is reported; ** means p ≤ 0.01; * means p ≤ 0.05; n.s. means p ≥ 0.05; sig. NR = not reported if the difference is significant.

Table 4 Overview of the impact of AI-DSS use on the task performance of experts, listed by publication date (continued)

Study	*Detailed sample description*	*Experimental task*	*AI-DSS (in the experimental condition)*	*Decision quality in terms of*					*Decision efficiency in terms of*	
				Accuracy	*AUC*	*Sensitivity*	*Specificity*	*Other criteria*	*Average processing time*	*Other criteria*
Calisto et al. (2022)	N = 45 physicians (seniors, middles, juniors and interns)	Medical diagnosis (breast cancer)	System offers several functions, e.g., diagnosis advice, possibility to visualise and manipulate images.			↑ (sig. NR)		False negative rate: ↓ (sig. NR) False positive rate: ↓ (sig. NR) Precision: ↑ (sig. NR) Recall: ↑ (sig. NR)	↓ (sig. NR)	
Duchevet et al. (2022)	N = 7 pilots	Go-arounds during the final approach	System alarms in case of parameter deviations, e.g., at the stabilisation gate if a go-around is assumed to be needed.					Stabilisation: ↓(n.s.)		
Finck et al. (2022)	N = 4 neuroradiologists (N = 2 experienced, N = 2 inexperienced)	Medical diagnosis (of head CT scans as 'normal' or 'pathological')	System anomalies for head computed tomography (CT), tailored to provide patient-level triage and voxel-based highlighting of pathologies.	↑**				False positive rate: ↓**	↓**	
Henkel et al. (2022)	N = 10 urologists	Medical treatment decisions (prostate cancer)	System provides recommendations on diagnostic and therapeutic options.						↓**	

Notes: AUC: area under the receiver operating characteristic curve; ↑ = increase; ↓ = decrease; → = stable; no arrow means no direction is reported; ** means p ≤ 0.01; * means p ≤ 0.05; n.s. means p ≥ 0.05; sig. NR = not reported if the difference is significant.

Table 4 Overview of the impact of AI-DSS use on the task performance of experts, listed by publication date (continued)

Study	*Detailed sample description*	*Experimental task*	*AI-DSS (in the experimental condition)*	*Decision quality in terms of*					*Decision efficiency in terms of*	
				Accuracy	*AUC*	*Sensitivity*	*Specificity*	*Other criteria*	*Average processing time*	*Other criteria*
Hwang et al. (2022)	N = 9 polysomnographic technicians	Scoring sleep stages	System provides sleep staging predictions and related sound explanations.					Macro-F1 scores: ↑**		
Jussupow et al. (2021)	N = 47 medical students (N = 26 without clinical experience; N = 21 with medical experience)	Medical diagnosis (chronic obstructive pulmonary disease, short COPD, vs. no COPD)	System predicts pulmonary function values from a CT scan for diagnosing COPD (50% accuracy).	↓ (sig. NR)						
Kiefer et al. (2022)	N = 7 physicians	Medical diagnosis (*Cornea guttata* in microscope images)	System predicts the diagnostic probability and also has other features, such as the ability to compare the image to be diagnosed with similar images from past diagnoses.			↑ (sig. NR)	↓ (sig. NR)	F-score: ↑ (sig. NR) Precision: ↑ (sig. NR)	↑ (sig. NR)	
Lacroux and Martin-Lacroux (2022)	N = 694 experienced person in personnel selection	Personnel selection	System ranks resumes according to a given job description (50% accuracy).	↓ (n.s.)						

Notes: AUC: area under the receiver operating characteristic curve; ↑ = increase; ↓ = decrease; → = stable; no arrow means no direction is reported; ** means p ≤ 0.01; * means p ≤ 0.05; n.s. means p ≥ 0.05; sig. NR = not reported if the difference is significant.

Table 4 Overview of the impact of AI-DSS use on the task performance of experts, listed by publication date (continued)

Study	*Detailed sample description*	*Experimental task*	*AI-DSS (in the experimental condition)*	*Decision quality in terms of*					*Decision efficiency in terms of*	
				Accuracy	*AUC*	*Sensitivity*	*Specificity*	*Other criteria*	*Average processing time*	*Other criteria*
Roller et al. (2022)	N = 8 physicians (N = 4 juniors, N = 4 seniors)	Medical decision (how high the risk of rejection and graft failure resulting in death is after a kidney transplant)	System for recognising patients at risk of rejection and death-censored graft failure (rejection prediction: 0.747 AUC, 56% sensitivity, 69% specificity; graft loss prediction: 0.964 AUC; 67% sensitivity; 92% specificity).		Task 1 rejection: senior subjects ↓ (sig. NR); junior subjects ↑ (sig. NR)	Task 1 rejection: all subjects ↓ (sig. NR)	Task 1 rejection: all subjects ↑ (sig. NR)			
					Task 2 graft los: senior subjects ↓ (sig. NR); junior subjects ↑ (sig. NR)	Task 2 graft los: all subjects → (sig. NR)	Task 2 graft los: all subjects ↑ (sig. NR)			
Kim et al. (2023)	N = 49 nurses	Pressure ulcer (PU) prevention	System provides support by prevention PU, e.g., in the form of risk prediction.					Perceived degree of nursing performance: ↑*		
Kindler et al. (2023)	N = 2 senior pathologists	Medical diagnosis (lymph node metastases)	System creates a colour-coded heat map for possible cancer areas on slides (95.2–99.6% accuracy).							Median processing time: ↓**

Notes: AUC: area under the receiver operating characteristic curve; ↑ = increase; ↓ = decrease; → = stable; no arrow means no direction is reported; ** means p ≤ 0.01; * means p ≤ 0.05; n.s. means p ≥ 0.05; sig. NR = not reported if the difference is significant.

Table 4 Overview of the impact of AI-DSS use on the task performance of experts, listed by publication date (continued)

Study	*Detailed sample description*	*Experimental task*	*AI-DSS (in the experimental condition)*	*Decision quality in terms of*					*Decision efficiency in terms of*	
				Accuracy	*AUC*	*Sensitivity*	*Specificity*	*Other criteria*	*Average processing time*	*Other criteria*
Laursen et al. (2023)	N = 13 physicians form various departments in which haemorrhages are relevant	Identification of haemorrhage during the review of an admission chart review	System highlights phrases in text (such as chart reviews) that indicate haemorrhage (93.7% sensitivity and 98.1% specificity).	↑ (sig. NR)						
Lee and Chew (2023)	N = 7 therapists	Rehabilitation assessment	System predicts the quality of motion of a patient based on the rehabilitation exercises of the patient based on video data and generate user-specific analysis that includes feature analysis, images of salient frames, and graphs of joint trajectories. The system exists in two different forms, one with salient feature explanations and one with counterfactual explanations.					F-score for system with salient features explanations: ↓** F-score for system with counter-factual explanations: ↑ (n.s.)		

Notes: AUC: area under the receiver operating characteristic curve; ↑ = increase; ↓ = decrease; → = stable; no arrow means no direction is reported; ** means p ≤ 0.01; * means p ≤ 0.05; n.s. means p ≥ 0.05; sig. NR = not reported if the difference is significant.

Table 4 Overview of the impact of AI-DSS use on the task performance of experts, listed by publication date (continued)

Study	*Detailed sample description*	*Experimental task*	*AI-DSS (in the experimental condition)*	*Decision quality in terms of*					*Decision efficiency in terms of*	
				Accuracy	*AUC*	*Sensitivity*	*Specificity*	*Other criteria*	*Average processing time*	*Other criteria*
Shah et al. (2023)	N = 10 radiology trainees	Medical diagnosis based on brain MRIs	System displays probabilities for various diagnoses, but only after the user had feet them with some information beforehand.						↑ (n.s.)	
Yoon et al. (2023)	N = 66 ophthalmologists (N = 36 retina experts, N = 30 non-retina experts)	Medical diagnosis of retinal diseases using optical coherence tomography	System displays probabilities for various retinal disease (96.3% accuracy, 97.1% sensitivity, 95.7% specificity; 98.4% AUC).	Retina experts: ↑ (sig. NR) Non-retina experts: ↑ (sig. NR)	Retina experts: ↑** Non-retina experts: ↑**					
Zhang et al. (2023)	N = 2 radiologists	Medical diagnosis (COVID-19 vs. common pneumonia)	System for detecting COVID-19 diseases by screening, assessing, and segmenting lesions.	↑ (sig. NR)					↓ (sig. NR)	

Notes: AUC: area under the receiver operating characteristic curve; ↑ = increase; ↓ = decrease; → = stable; no arrow means no direction is reported; ** means $p \leq 0.01$; * means $p \leq 0.05$; n.s. means $p \geq 0.05$; sig. NR = not reported if the difference is significant.

Table 5 Detailed overview of the effects of AI-DSS usage on task accuracy of experts, listed by effect size

Study	*Sample*	*Detailed description of the experimental task (in AI and non-AI conditions[1])*		*AI-DSS task accuracy in %*	*Experts' task accuracy (± SD[2]) in % without AI*	*Experts' task accuracy (± SD[2]) in % with AI*	*Significant difference in expert accuracy with vs. without AI*	*Self-calculated Cohen's d to quantify the degree of change in task accuracy with AI support[3]*
Laursen et al. (2023)[4]	N = 7 physicians	Review a patient record from admission and find all haemorrhages (63 can be found)		NR	45 (±8)	93 (±13)	NR	d = 4.45
	N = 6 physicians	Review a patient record from admission and find all haemorrhages (51 can be found)		NR	26 (±17)	75 (±10)	NR	d = 3.51
Didimo et al. (2018)[4]	N = 32 employees in a fiscal audit office	Two sets of tasks that require the identification of certain data of a specific taxpayer, such as the fiscal code of its shareholders	Task set 2	NR	63.09 (±NR)	98.83 (±NR)	**	d = 3.27
			Task set 1	NR	87.08 (±NR)	98.98 (±NR)	**	d = 1.59
Yoon et al. (2023)[4]	N = 36 ophthalmologists who are retina experts	Review a total of 100 OCT images to determine whether they show acute or chronic eye disease		96.3	91.8 (±4.29)	95.2 (±3.67)	NR	d = 0.85

Notes: [1] The experimental task described applies to both the baseline condition (without AI support) and the intervention condition (with AI support), unless explicitly stated otherwise.

[2] In most studies, the SDs are not reported directly, but rather, the confidence intervals. In these cases, we calculated the SDs from the confidence intervals and the sample size.

[3] We calculated Cohen's d for all studies ourselves. Specifically, we either:

a converted the odds ratios reported in the studies into Cohen's d (Kiani et al., 2020; Lacroux and Martin-Lacroux, 2022)

b converted the Mann-Whitney U values reported in the studies into Cohen's d (Didimo et al., 2018)

c directly calculated Cohen's d from the means, SDs, and sample sizes provided in the studies (Bai et al., 2020; Jacobs et al., 2021; Laursen et al., 2023; Rajpurkar et al., 2020; Yang et al., 2021; Yoon et al., 2023)

d when the SD was not reported in the study, we estimated an approximation of Cohen's d by first calculating the chi-square value from a 2 × 2 contingency table. Using this value and the total number of observations, we derived Cramér's V, which was then converted into Cohen's d (Finck et al., 2022; Jussupow et al., 2021; Kavya et al., 2021; Nam et al., 2021; Rudie et al., 2021; Zhang et al., 2023).

[4] Studies only report the results separately for different tasks and/or different samples.

[5] The Cohen's d reported by Kiani et al. (2020) is from a statistical analysis controlling for experts' experience levels and task difficulty.

SD = standard deviation; NR = not reported; ** means $p \leq 0.01$; * means $p \leq 0.05$; n.s. means $p \geq 0.05$; CT = computed tomography; OCT = optical coherence tomography; MRI = magnetic resonance imaging; WSI = whole-slide image; COPD = chronic obstructive pulmonary disease.

Table 5 Detailed overview of the effects of AI-DSS usage on task accuracy of experts, listed by effect size (continued)

Study	*Sample*	*Detailed description of the experimental task (in AI and non-AI conditions[1])*	*AI-DSS task accuracy in %*	*Experts' task accuracy (± SD[2]) in % without AI*	*Experts' task accuracy (± SD[2]) in % with AI*	*Significant difference in expert accuracy with vs. without AI*	*Self-calculated Cohen's d to quantify the degree of change in task accuracy with AI support[3]*
Yoon et al. (2023)[4]	N = 30 ophthalmologists who are non retina experts	Review a total of 100 OCT images to determine whether they show acute or chronic eye disease	96.3	85.9 (±5.31)	90.5 (±5.87)	NR	d = 0.82
Bai et al. (2020)	N = 6 radiologists	Review a total of 119 chest CT images to determine whether they show COVID-19 or pneumonia from other causes	87	85 (±8.13)	90 (±6.88)	**	d = 0.66
Rajpurkar et al. (2020)	N = 13 physicians	Review a total of 114 patient records and chest X-rays to determine whether or not the patients have tuberculosis (note: subjects process half of the cases with AI and half without AI)	79	60 (±5.52)	65 (±9.21)	*	d = 0.66

Notes: [1] The experimental task described applies to both the baseline condition (without AI support) and the intervention condition (with AI support), unless explicitly stated otherwise.

[2] In most studies, the SDs are not reported directly, but rather, the confidence intervals. In these cases, we calculated the SDs from the confidence intervals and the sample size.

[3] We calculated Cohen's d for all studies ourselves. Specifically, we either:

a converted the odds ratios reported in the studies into Cohen's d (Kiani et al., 2020; Lacroux and Martin-Lacroux, 2022)

b converted the Mann-Whitney U values reported in the studies into Cohen's d (Didimo et al., 2018)

c directly calculated Cohen's d from the means, SDs, and sample sizes provided in the studies (Bai et al., 2020; Jacobs et al., 2021; Laursen et al., 2023; Rajpurkar et al., 2020; Yang et al., 2021; Yoon et al., 2023)

d when the SD was not reported in the study, we estimated an approximation of Cohen's d by first calculating the chi-square value from a 2 × 2 contingency table. Using this value and the total number of observations, we derived Cramér's V, which was then converted into Cohen's d (Finck et al., 2022; Jussupow et al., 2021; Kavya et al., 2021; Nam et al., 2021; Rudie et al., 2021; Zhang et al., 2023).

[4] Studies only report the results separately for different tasks and/or different samples.

[5] The Cohen's d reported by Kiani et al. (2020) is from a statistical analysis controlling for experts' experience levels and task difficulty.

SD = standard deviation; NR = not reported; ** means $p \leq 0.01$; * means $p \leq 0.05$; n.s. means $p \geq 0.05$; CT = computed tomography; OCT = optical coherence tomography; MRI = magnetic resonance imaging; WSI = whole-slide image; COPD = chronic obstructive pulmonary disease.

Table 5 Detailed overview of the effects of AI-DSS usage on task accuracy of experts, listed by effect size (continued)

Study	*Sample*	*Detailed description of the experimental task (in AI and non-AI conditions[1])*	*AI-DSS task accuracy in %*	*Experts' task accuracy (± SD[2]) in % without AI*	*Experts' task accuracy (± SD[2]) in % with AI*	*Significant difference in expert accuracy with vs. without AI*	*Self-calculated Cohen's d to quantify the degree of change in task accuracy with AI support[3]*
Rudie et al. (2021)[4]	N = 4 radiology resident	Review a total of 194 MRIs of the brain to determine which type of brain disease is present (note: subjects process half of the cases with AI and half without AI)	61	30 (±NR)	55 (±NR)	**	d = 0.52
Zhang et al. (2023)	N = 2 radiologists	Review a total of 116 chest CT images to determine whether they show COVID-19 or common pneumonia	NR	83.7 (±NR)	97.7 (±NR)	n.s.	d = 0.51
Yang et al. (2021)	N = 3 radioloigists	Review a total of 185 chest CT images to determine whether they show COVID-19 or non-COVID patients	91.4	94.1 (±3.18)	95.1 (±2.96)	-	d = 0.33

Notes: [1] The experimental task described applies to both the baseline condition (without AI support) and the intervention condition (with AI support), unless explicitly stated otherwise.

[2] In most studies, the SDs are not reported directly, but rather, the confidence intervals. In these cases, we calculated the SDs from the confidence intervals and the sample size.

[3] We calculated Cohen's d for all studies ourselves. Specifically, we either:

a converted the odds ratios reported in the studies into Cohen's d (Kiani et al., 2020; Lacroux and Martin-Lacroux, 2022)

b converted the Mann-Whitney U values reported in the studies into Cohen's d (Didimo et al., 2018)

c directly calculated Cohen's d from the means, SDs, and sample sizes provided in the studies (Bai et al., 2020; Jacobs et al., 2021; Laursen et al., 2023; Rajpurkar et al., 2020; Yang et al., 2021; Yoon et al., 2023)

d when the SD was not reported in the study, we estimated an approximation of Cohen's d by first calculating the chi-square value from a 2 × 2 contingency table. Using this value and the total number of observations, we derived Cramér's V, which was then converted into Cohen's d (Finck et al., 2022; Jussupow et al., 2021; Kavya et al., 2021; Nam et al., 2021; Rudie et al., 2021; Zhang et al., 2023).

[4] Studies only report the results separately for different tasks and/or different samples.

[5] The Cohen's d reported by Kiani et al. (2020) is from a statistical analysis controlling for experts' experience levels and task difficulty.

SD = standard deviation; NR = not reported; ** means $p \leq 0.01$; * means $p \leq 0.05$; n.s. means $p \geq 0.05$; CT = computed tomography; OCT = optical coherence tomography; MRI = magnetic resonance imaging; WSI = whole-slide image; COPD = chronic obstructive pulmonary disease.

Table 5 Detailed overview of the effects of AI-DSS usage on task accuracy of experts, listed by effect size (continued)

Study	*Sample*	*Detailed description of the experimental task (in AI and non-AI conditions[1])*	*AI-DSS task accuracy in %*	*Experts' task accuracy (± SD[2]) in % without AI*	*Experts' task accuracy (± SD[2]) in % with AI*	*Significant difference in expert accuracy with vs. without AI*	*Self-calculated Cohen's d to quantify the degree of change in task accuracy with AI support[3]*
Finck et al. (2022)	N = 4 neuroradiologists	Review a total of 80 head CT scans to determine whether they show 'normal' or 'pathological' brains	NR	96.6 (±NR)	99.1 (±NR)	**	d = 0.18
Kiani et al. (2020)	N = 11 pathologists	Review a total of 80 WSIs to determine whether they show hepatocellular carcinoma or cholangiocarcinoma liver cancer	84.2	89.8 (±3.47)	91.4 (±3.13)	n.s.	d = 0.145
Kavya et al. (2021)	N = 4 physicians	Review a total of 169 patient records to determine their allergy diagnosis	86.39	77.2 (±NR)	81.8 (±NR)	-	d = 0.12
Nam et al. (2021)	N = 6 radiologists	Review a total of 202 chest X-rays and write a short formal report, including abnormal findings and possible differential diagnoses	NR	86.8[4] (±NR)	90.5[4] (±NR)	*	d = 0.12

Notes: [1] The experimental task described applies to both the baseline condition (without AI support) and the intervention condition (with AI support), unless explicitly stated otherwise.

[2] In most studies, the SDs are not reported directly, but rather, the confidence intervals. In these cases, we calculated the SDs from the confidence intervals and the sample size.

[3] We calculated Cohen's d for all studies ourselves. Specifically, we either:

a converted the odds ratios reported in the studies into Cohen's d (Kiani et al., 2020; Lacroux and Martin-Lacroux, 2022)

b converted the Mann-Whitney U values reported in the studies into Cohen's d (Didimo et al., 2018)

c directly calculated Cohen's d from the means, SDs, and sample sizes provided in the studies (Bai et al., 2020; Jacobs et al., 2021; Laursen et al., 2023; Rajpurkar et al., 2020; Yang et al., 2021; Yoon et al., 2023)

d when the SD was not reported in the study, we estimated an approximation of Cohen's d by first calculating the chi-square value from a 2 × 2 contingency table. Using this value and the total number of observations, we derived Cramér's V, which was then converted into Cohen's d (Finck et al., 2022; Jussupow et al., 2021; Kavya et al., 2021; Nam et al., 2021; Rudie et al., 2021; Zhang et al., 2023).

[4] Studies only report the results separately for different tasks and/or different samples.

[5] The Cohen's d reported by Kiani et al. (2020) is from a statistical analysis controlling for experts' experience levels and task difficulty.

SD = standard deviation; NR = not reported; ** means $p \leq 0.01$; * means $p \leq 0.05$; n.s. means $p \geq 0.05$; CT = computed tomography; OCT = optical coherence tomography; MRI = magnetic resonance imaging; WSI = whole-slide image; COPD = chronic obstructive pulmonary disease.

Table 5 Detailed overview of the effects of AI-DSS usage on task accuracy of experts, listed by effect size (continued)

Study	*Sample*	*Detailed description of the experimental task (in AI and non-AI conditions[1])*	*AI-DSS task accuracy in %*	*Experts' task accuracy (± SD[2]) in % without AI*	*Experts' task accuracy (± SD[2]) in % with AI*	*Significant difference in expert accuracy with vs. without AI*	*Self-calculated Cohen's d to quantify the degree of change in task accuracy with AI support[3]*
Rudie et al. (2021)[4]	N = 3 neuroradiologists	Review a total of 194 MRIs of the brain to determine which type of brain disease is present (note: subjects process half of the cases with AI and half without AI)	61	69 (±NR)	72 (±NR)	n.s.	d = 0.06
Jacobs et al. (2021)	N = 220 physicians	Review a total of 17 patient information records and select their antidepressant treatment (note: subjects process 5 cases without AI and 12 cases with AI)	66.7	35.7 (±18.1)	35.6 (±11.7)	n.s.	d = –0.01
Jussupow et al. (2021)	N = 47 medical students	Review a total of three lung CT scans to determine whether they show COPD or no-COPD (note: subjects process one case without AI and two cases with AI)	50	77 (±NR)	72.09 (±NR)	NR	d = –0.10

Notes: [1] The experimental task described applies to both the baseline condition (without AI support) and the intervention condition (with AI support), unless explicitly stated otherwise.

[2] In most studies, the SDs are not reported directly, but rather, the confidence intervals. In these cases, we calculated the SDs from the confidence intervals and the sample size.

[3] We calculated Cohen's d for all studies ourselves. Specifically, we either:

a converted the odds ratios reported in the studies into Cohen's d (Kiani et al., 2020; Lacroux and Martin-Lacroux, 2022)

b converted the Mann-Whitney U values reported in the studies into Cohen's d (Didimo et al., 2018)

c directly calculated Cohen's d from the means, SDs, and sample sizes provided in the studies (Bai et al., 2020; Jacobs et al., 2021; Laursen et al., 2023; Rajpurkar et al., 2020; Yang et al., 2021; Yoon et al., 2023)

d when the SD was not reported in the study, we estimated an approximation of Cohen's d by first calculating the chi-square value from a 2 × 2 contingency table. Using this value and the total number of observations, we derived Cramér's V, which was then converted into Cohen's d (Finck et al., 2022; Jussupow et al., 2021; Kavya et al., 2021; Nam et al., 2021; Rudie et al., 2021; Zhang et al., 2023).

[4] Studies only report the results separately for different tasks and/or different samples.

[5] The Cohen's d reported by Kiani et al. (2020) is from a statistical analysis controlling for experts' experience levels and task difficulty.

SD = standard deviation; NR = not reported; ** means $p \leq 0.01$; * means $p \leq 0.05$; n.s. means $p \geq 0.05$; CT = computed tomography; OCT = optical coherence tomography; MRI = magnetic resonance imaging; WSI = whole-slide image; COPD = chronic obstructive pulmonary disease.

Table 5 Detailed overview of the effects of AI-DSS usage on task accuracy of experts, listed by effect size (continued)

Study	*Sample*	*Detailed description of the experimental task (in AI and non-AI conditions[1])*	*AI-DSS task accuracy in %*	*Experts' task accuracy (± SD[2]) in % without AI*		*Experts' task accuracy (± SD[2]) in % with AI*		*Significant difference in expert accuracy with vs. without AI*	*Self-calculated Cohen's d to quantify the degree of change in task accuracy with AI support[3]*
Lacroux and Martin-Lacroux (2022)	N = 694 experienced person in personnel selection	Analyse a job description for an HR manager and two resume abstracts (unequally qualified) to identify the more suitable candidate	50	64.2	(±NR)	56.1	(±NR)	n.s.	d = –0.19
Lacroux and Martin-Lacroux (2022)	N = 694 experienced person in personnel selection	Analyse a job description for an HR manager and two resume abstracts (unequally qualified) to identify the more suitable candidate	50	64.2	(±NR)	56.1	(±NR)	n.s.	d = –0.19

Notes: [1] The experimental task described applies to both the baseline condition (without AI support) and the intervention condition (with AI support), unless explicitly stated otherwise.

[2] In most studies, the SDs are not reported directly, but rather, the confidence intervals. In these cases, we calculated the SDs from the confidence intervals and the sample size.

[3] We calculated Cohen's d for all studies ourselves. Specifically, we either:

a converted the odds ratios reported in the studies into Cohen's d (Kiani et al., 2020; Lacroux and Martin-Lacroux, 2022)

b converted the Mann-Whitney U values reported in the studies into Cohen's d (Didimo et al., 2018)

c directly calculated Cohen's d from the means, SDs, and sample sizes provided in the studies (Bai et al., 2020; Jacobs et al., 2021; Laursen et al., 2023; Rajpurkar et al., 2020; Yang et al., 2021; Yoon et al., 2023)

d when the SD was not reported in the study, we estimated an approximation of Cohen's d by first calculating the chi-square value from a 2 × 2 contingency table. Using this value and the total number of observations, we derived Cramér's V, which was then converted into Cohen's d (Finck et al., 2022; Jussupow et al., 2021; Kavya et al., 2021; Nam et al., 2021; Rudie et al., 2021; Zhang et al., 2023).

[4] Studies only report the results separately for different tasks and/or different samples.

[5] The Cohen's d reported by Kiani et al. (2020) is from a statistical analysis controlling for experts' experience levels and task difficulty.

SD = standard deviation; NR = not reported; ** means $p \leq 0.01$; * means $p \leq 0.05$; n.s. means $p \geq 0.05$; CT = computed tomography; OCT = optical coherence tomography; MRI = magnetic resonance imaging; WSI = whole-slide image; COPD = chronic obstructive pulmonary disease.

Table 6 Detailed overview of the effects of AI-DSS usage on the task processing time of experts, listed by effect size

Study	*Sample*	*Detailed description of the experimental task (in AI and non-AI condition[1])*		*Experts' task processing time (± SD) in seconds without AI*		*Experts' task processing time (± SD) in seconds with AI*		*Significant difference in experts' task processing time with vs. without AI*	*Self-calculated Cohen's d to quantify the degree of change in task processing time with AI support[2]*
Kim et al. (2020)	N = 3 physicians	Review a total of 387 chest radiographs and grade all of the radiographs on a five-point scale for the presence of pneumonia (as follows: 1 = definitely normal, 2 = probably normal, 3 = indeterminate, 4 = probably pneumonia, and 5 = definitely pneumonia)		9,900 total	(±737.4)	6,060 total	(±982.2)	NR	d = 4.42
Didimo et al. (2018)	N = 32 employees in a fiscal audit office	Two sets of tasks that require the identification of certain data of a specific taxpayer, such as the fiscal code of its shareholders	Task set 1	341.17 total	(±NR)	178.48 total	(±NR)	**	d = 3.01
			Task set 2	435.81 total	(±NR)	159.92 total	(±NR)	**	d = 3.01
Calisto et al. (2022)	N = 45 physicians	Review a total of three multimodality patient images (i.e., mammography, ultrasound and magnetic resonance) and determine the risk of breast cancer		377 total	(±44.56)	308 total	(±57.03)	NR	d = 1.35
Finck et al. (2022)	N = 4 neuroradiologists	Review a total of 80 head CT scans to determine whether they show 'normal' or 'pathological' brains		65.1 case	(±8.9)	54.9 case	(±7.1)	**	d = 1.27
Sung et al. (2021)	N = 6 radiologists	Review a total of 228 chest radiographs and detect and localise major abnormal findings such as nodules and consolidation (note: subjects process half of the cases with AI and half without AI)		24 case	(±21)	12 case	(±8)	**	d = 0.76

Notes: [1]The experimental task described applies to both the baseline condition (without AI support) and the intervention condition (with AI support), unless explicitly stated otherwise. [2]We calculated Cohen's d ourselves using the mean, SDs, and sample size, except for the study of Didimo et al. (2018), where we transformed the Mann-Whitney U value into Cohen's. Total = the average processing time refers to the mean time required to process the entire set of tasks; case = the average processing time refers to the mean time required to process a case within the task set; SD = standard deviation; NR = not reported; ** means p ≤ 0.01; * means p ≤ 0.05; n.s. means p ≥ 0.05; CT = computed tomography; MRI = magnetic resonance imaging.

Table 6 Detailed overview of the effects of AI-DSS usage on the task processing time of experts, listed by effect size (continued)

Study	*Sample*	*Detailed description of the experimental task (in AI and non-AI condition[1])*	*Experts' task processing time (± SD) in seconds without AI*	*Experts' task processing time (± SD) in seconds with AI*	*Significant difference in experts' task processing time with vs. without AI*	*Self-calculated Cohen's d to quantify the degree of change in task processing time with AI support[2]*
Kozuka et al. (2020)	N = 2 radiologists	Review a total of 120 chest CT images of patients with suspected lung cancer, marking each nodule and annotating the type of nodule	373 total (±66.47)	331 total (±55.15)	NR	d = 0.67
Martini et al. (2021)	N = 6 radiologists	Review a total of 100 chest CT images of consecutive oncology patients and mark the presence, size and location of lung nodules	194 case (±126)	154 case (±134)	**	d = 0.31
Nam et al. (2021)	N = 6 radiologists	Review a total of 202 chest X-rays and write a short formal report, including abnormal findings and possible differential diagnoses	23.5 case (±23.7)	20.5 case (±22.8)	**	d = 0.13
Shah et al. (2023)	N = 10 radiology trainees	Review a total of 50 brain MRIs from a routine clinical worklist and make a differential diagnosis (note: subjects process half of the cases with AI and half without AI)	18.73 case (±10.8)	20.31 case (±11.19)	n.s.	d = –0.14

Notes: [1]The experimental task described applies to both the baseline condition (without AI support) and the intervention condition (with AI support), unless explicitly stated otherwise. [2]We calculated Cohen's d ourselves using the mean, SDs, and sample size, except for the study of Didimo et al. (2018), where we transformed the Mann-Whitney U value into Cohen's. Total = the average processing time refers to the mean time required to process the entire set of tasks; case = the average processing time refers to the mean time required to process a case within the task set; SD = standard deviation; NR = not reported; ** means $p \leq 0.01$; * means $p \leq 0.05$; n.s. means $p \geq 0.05$; CT = computed tomography; MRI = magnetic resonance imaging.

3.2.2 Experts' AI-performance compared to the individual performance of AI-DSS without expert validation

Of the 38 studies that address the influence of AI support on expert performance, in only ten studies, the expert performance with and without AI-DSS and individual performance of the AI-DSS were reported. This allowed a comparison of all three values across various studies (see Table 5), and the following trends were observed:

1. when the individual performances of humans and AI were comparable and relatively high, their combined performance outperformed that of each individual actor (Bai et al., 2020; Kiani et al., 2020; Yang et al., 2021)
2. when the experts performance was significantly below the AI performance, the human-AI interaction continued to boost human performance without outperforming the individual AI actor performance (Kavya et al., 2021; Rajpurkar et al., 2020; Rudie et al., 2021; Yoon et al., 2023)
3. when the AI performance was below human performance at a moderate level, the human-AI performance result was below individual human performance (Jussupow et al., 2021; Lacroux and Martin-Lacroux, 2022).

3.2.3 Experts' psychological load experience in terms of workload

Of the included studies, three examined the impact of an AI-DSS on the workloads experienced by experts during decision making (Cai et al., 2019; Duchevet et al., 2022; Lee et al., 2020). The authors asked their participants to rate their perceived mental effort, frustration (Cai et al., 2019; Lee et al., 2020), and cognitive workload (Duchevet et al., 2022) immediately after completing the experimental tasks (with vs. without AI support) using standardized questions. Both Cai et al. (2019) and Lee et al. (2020) observed that participants supported by an AI-DSS perceived significantly less mental effort during task completion than those supported by a traditional information system. However, neither study found any significant differences in perceived frustration. Duchevet et al. (2022) also found no significant difference in the cognitive workload reported by participants between the experimental and control conditions (with and without AI support). However, the authors reported that in the debriefings, experts indicated that the AI support freed up cognitive resources to focus on things that were important to them.

3.2.4 Experts' psychological load experience in terms of decision confidence and perceived safety

All three studies that evaluated the effect of AI-DSS on decision confidence found that individual confidence of experts in their decisions was similar with and without AI support (Finck et al., 2022; Shah et al., 2023; Zhou et al., 2021), despite two studies simultaneously reporting that experts made significantly better decisions (Finck et al., 2022; Zhou et al., 2021). Consistent with these findings, Duchevet et al. (2022) found that safety perception of pilots during simulated operations was approximately the same with and without AI support.

3.3 Results for RQ2: role of accuracy and explainability of AI-DSS on load experienced by and performance behaviour of experts in decision-making situations

3.3.1 Role of system accuracy for experts' task performance

Contrary to what may be initially assumed, none of the included studies (Table 7) examined the effects of different accuracy levels of AI-based systems on user experience and behaviour in AI-based situations. Instead, these studies fundamentally examined user responses to correct versus incorrect AI-DSS system advice in decision-making situations.

All five studies confirmed that the correctness of a system advice significantly influenced a user task performance. Specifically, all the studies showed that the decision quality of experts was significantly lower when they received incorrect advice than when they received correct advice. This significant deterioration in their performance owing to incorrect advice was also evident when compared with the baseline condition (no AI system support) (Table 7). However, the negative impacts of incorrect AI advice compared to no AI advice varied in magnitude across the studies. Our calculations showed that, according to Cohen (1988), the values ranged from no effect (d = –0.16, Jacobs et al., 2021) to a small effect (d = –0.35, Lacroux and Martin-Lacroux, 2022; d = –0.43, Jussupow et al., 2021) to a medium effect (d = –0.76, Kiani et al., 2020).

By contrast, the comparison between no advice and correct advice showed no significant differences in most studies. Only one of the four studies (Table 7) reported a significant improvement in performance with correct advice compared with no advice. In their study, Kiani et al. (2020) differed somewhat in their study design, particularly in the distribution of correct and incorrect advice. In the within-subjects design, the subjects received correct and incorrect advice on approximately 84 and 16 % of the AI-assisted decisions to be made, respectively. Jussupow et al. (2021) and Jacobs et al. (2021) also used a within-subject design in their work; however, the subjects in Jussupow et al.'s (2021) research received correct advice in only 50% of the AI-assisted cases, and in Jacobs et al.'s (2021) work they received correct advice in approximately 67% (eight out of 12) of the cases. Lacroux and Martin-Lacroux (2022) used a between-subject design, and their subjects solved only one case; whereas, the subjects in the study by Kiani et al. (2020) solved 160 cases (80 with and 80 without AI support).

3.3.2 Role of system accuracy for experts' decision confidence

Jacobs et al. (2021) and Gaube et al. (2021) examined the effect of the correctness of a system advice on the confidence of physicians in their decisions. However, these studies observed different effects. Gaube et al. (2021) observed that all participants, including both high- and low-level experts, were significantly more confident in their diagnoses when the advice was accurate. By contrast, Jacobs et al. (2021) found no significant effect of the correctness of advice on physicians' confidence in medical treatment decisions among conditions without system support, correct AI-based advice, and incorrect AI-based advice.

Table 7 Overview of studies that examined the influence of the correctness of a system recommendation on users

					Experimental condition			*Baseline condition (no AI support): Number of decision cases*	*Outcome*	
Authors	*Subjects*	*Task*	*AI-DSS under examination*	*Study design*	*Number of decision cases*	*Correct condition and frequency of the condition (within design) / number of group members (between design)*	*Incorrect condition and frequency of the condition (within design) / number of group members (between design)*		*Experts' decision quality*	*Experts' confidence with decision made*
Kiani et al. (2020)	N = 11 pathologists	Medical diagnosis	Systems help distinguish between the two most common types of primary liver cancer by giving a probability for each diagnosis, with an accompanying class activation map to help with interpretation	Within-subjects	160	Correct diagnostic advice (in 84% of the 80 AI-assisted cases)	Incorrect diagnostic advice (in 16% of the 80 AI-assisted cases)	80	correct > baseline**; incorrect < baseline**	-
Jacobs et al. (2021)	N = 220 physicians	Medical treatment choice	System gives a top-5 list of treatment advices (including a top-1 advice) for major depressive disorder	Within-subjects	17	Top-5 list of treatment advice including a correct top-1 advice (in 67% of 12 the AI-assisted cases)	Top 5 treatment advices including an incorrect top-1 advice (in 33% of 12 AI-assisted the cases)	5	incorrect < correct**; incorrect < baseline*; correct > baseline (n.s.)	incorrect < correct (n.s.); incorrect < baseline (n.s.); correct < baseline (n.s.)
Gaube et al. (2021)′	N = 265 physicians	Medical diagnosis	System analyses X-ray images of breasts and provides concrete information about abnormalities and a diagnostic advice on this basis	Within-subjects	8	Correct diagnostic advice (in 75% of the 8 AI-assisted cases)	Incorrect diagnostic advice (in 25% of the 8 AI-assisted cases)	-	incorrect < correct**	incorrect < correct**

Notes: ** = p < 0.01; * = p < 0.05; n.s = p > 0.05; – means not investigated; ′this study was the only one that did not investigate how the test subjects behaved with correct or incorrect recommendations compared to no support (baseline condition).

Table 7 Overview of studies that examined the influence of the correctness of a system recommendation on users (continued)

					Experimental condition			*Baseline condition (no AI support): Number of decision cases*	*Outcome*	
Authors	*Subjects*	*Task*	*AI-DSS under examination*	*Study design*	*Number of decision cases*	*Correct condition and frequency of the condition (within design) / number of group members (between design)*	*Incorrect condition and frequency of the condition (within design) / number of group members (between design)*		*Experts' decision quality*	*Experts' confidence with decision made*
Jussupow et al. (2021)	N = 47 novice physicians	Medical diagnosis	System predicts pulmonary function values from a computed tomography (CT) scan for diagnosing COPD and gives advice based on the analysis of the data and diagnostic advice: COPD/NO COPD	Within-subjects	3	Correct diagnostic advice (in 50% of the 2 AI-assisted cases)	Incorrect diagnostic advice (in 50% of the 2 AI-assisted cases)	1	incorrect < correct**; incorrect < baseline*; correct > baseline (n.s.)	-
Lacroux and Martin-Lacroux (2022)	Approx. N = 416 recruiting experts	Personnel selection	System rank resumes to a given job description	Between-subjects	1	Recommendation of the most suitable candidate (received by approx. 139 participants)	Recommendation of a candidate who is not the most suitable (received by approx. 139 participants)	Approx. 139 participants	incorrect < correct **; incorrect < baseline **; correct > baseline (n.s.)	-

Notes: ** = p < 0.01; * = p < 0.05; n.s = p > 0.05; – means not investigated; ′this study was the only one that did not investigate how the test subjects behaved with correct or incorrect recommendations compared to no support (baseline condition).

3.3.3 Role of system explanations for experts' task performance

Six of the included studies focused on the effect of system explanations on users' task performance (Gaube et al., 2023; Hwang et al., 2022; Jacobs et al., 2021; Lee and Chew, 2023; Pushparaj et al., 2023; Yoon et al., 2023). To this end, four studies compared whether users react differently to AI-DSS when they receive, in addition to the system advice, its visual, textual, or auditory explanation. Pushparaj et al. (2023) found that experts could make faster decisions owing to additional explanations, with a large effect of d = 0.81. Gaube et al. (2023) and Hwang et al. (2022) observed that system explanations significantly increased the task performance of novices and non-task experts but did not affect that of experts. For instance, Gaube et al. (2023) found that non-task experts (physicians with internal or emergency medicine training) made better diagnostic decisions in reviewing radiographs when they received AI-based advice with visual annotations compared to when they did not receive such annotations; however, no significant effect was observed on the diagnostic accuracy of the radiology experts. Notably, the subjects' performance was considerably high throughout the experiment and only significantly lower in one case under both experimental conditions (with and without visual explanations). Interestingly, the annotations in this difficult case also appeared to affect the experts positively. The authors, therefore, assume that ceiling effects overlapped the effects in easier cases. Building on this, Yoon et al. (2023) noted that system explanations significantly improved performance for both non-experts and task experts. Interestingly, the performance enhancement effect was notably greater among non-experts compared to task experts.

Lee and Chew (2023) compared the effect of two different explanatory approaches. First, the widely used feature-based explanations, which "denotes how much each input feature contributes to a model's output for a given data point" [Bhatt et al., (2020), p.1] and are also used in the studies mentioned above and, second, counterfactual explanations.

The latter are 'what-if' explanations describing how inputs can be modified to achieve an AI outcome. The study showed that counterfactual explanations provided better support than salient feature explanations. Under this condition, the subjects made significantly more correct and fewer incorrect decisions. This is mainly because the experts in this condition showed less overtrust and rejected wrong AI outputs in 19% more cases than with salient-feature explanations. However, experts in the counterfactual condition were also 10% less likely to agree with the correct AI advice compared to the condition with salient feature explanations. This tendency towards critical thinking is also reflected in the processing time, which is on average 21 seconds longer for the system with counterfactual explanations compared to the one with salient feature analysis in a decision-making task. In their work, Jacobs et al. (2021) also compared whether experts react differently to correct and incorrect AI advice when given different types of explanations. Specifically, they compared the effect of feature-based and heuristic-based explanations. In this case, the results showed no significant differences in performance behaviour between the different explanatory approaches or the control condition without additional system explanations.

3.3.4 Role of system explanations for experts' decision confidence

Gaube et al. (2023), Panigutti et al. (2022) and Sivaraman et al. (2023) identified no differences in the assessment of the decision confidence of subjects with and without additional explanations of the system output. Jacobs et al. (2021) and Lee and Chew (2023) observed that different types of system explanations also had no differential effect on decision confidence.

3.3.5 Role of system explanations for experts' workload

Pushparaj et al. (2023), observed no difference in the self-perceived workload of the experts when they incorporated a system without or with additional explanations. However, the physiological data showed that the cognitive load was significantly higher in the condition with the additional explanations. Lee and Chew (2023) observed that the subjects perceived experimental tasks completed with an AI-DSS with feature-based explanations as less effortful ($p < 0.01$) and less frustrating ($p < 0.01$) than the tasks completed with an AI-DSS with counterfactuals explanations.

4 Discussion

In high-stakes work areas, employees often face high cognitive challenges, such as information overflow, complexity, and time and performance pressure, in their tasks (Walczok and Bipp, 2023). Physicians are required to make diagnoses or create treatment plans under enormous time pressure; bankers are required to create an optimum investment portfolio from various investment options for their clients. To assist their employees in such demanding decision-making situations, more and more organizations are planning to implement AI-DSS (Gartner, 2023; Naikar et al., 2023). With its support, decision-makers should feel relieved (e.g., less exhausted) during the decision-making process (Cai et al., 2019; Langer et al., 2021) and simultaneously act more effectively or efficiently, e.g., by making correct decisions more often (Finck et al., 2022; Langer et al., 2021). Ideally, the added value of human-AI decision-making is also demonstrated by the fact that the joint decision-making performance surpasses not only that of human experts but also the AI system alone (Bansal et al., 2021; Levy et al., 2021; Zhang et al., 2020).

To examine whether all these desired goals of using AI-DSS in the work context are achievable (RQ1) and which role accuracy and explainability of AI-DSS play in this context (RQ2), a systematic literature review was conducted. Thus identified 44 experimental studies investigated the effect of AI-DSS as a whole or the individual system characteristics on the psychological load experience and behaviour of experts in work-related decision-making situations.

4.1 Results examination

Notably, very few scholars have investigated the effects of AI support on the psychological load (RQ1b). For example, only three groups (Cai et al., 2019; Duchevet et al., 2022; Lee et al., 2020) examined whether AI-DSS usage reduces the perceived workload of experts in decision-making processes. Two of these three studies provided positive indications in this direction (Cai et al., 2019; Lee et al., 2020). However, the limited number of studies makes it difficult to draw general conclusions regarding RQ1b. Therefore, further investigation is required to address this question. In contrast, the question of whether and to what extent collaboration between humans and AI in work-related decision-making significantly enhances expert performance compared to their performance without an AI DSS has been frequently studied, allowing us to answer this aspect of RQ1a (see Subsection 3.2). First, the results show that AI support significantly reduces the processing time of experts, provided that the system does not have a complex interface. However, it is conceivable that positive effects may also occur with more extensive interfaces after a longer interaction phase, which should be investigated further in future studies (see Table 4). Second, in terms of the influence of AI on the accuracy with which users perform tasks, we found that a wide range of effects are possible. In some studies, AI significantly enhanced the performance of experts, while in others, the improvement was moderate or slight. In certain cases, AI had no impact on task performance, and in some instances, it even led to a decline in performance (see Table 5). The way in which AI-DSS affects expert performance appears to depend mainly on two factors. Firstly, the basic performance of the human experts without AI support and, secondly, the performance of the AI system itself.

First, the individual human performance logically determines the (potential) transformative power of an AI-DSS. If the human performance is already high because there is no or only a slight discrepancy between the expert resources and cognitive demands of the decision situation, for example, as observed in the study by Finck et al. (2022), the efficiency of the AI system is limited. Therefore, it is not surprising that Rudie et al. (2021), Calisto et al. (2022) and Didimo et al. (2018), who provided participants with tasks of varying difficulty, observed that the effect of AI support increased significantly with increasing decision complexity. Second, for AI-DSS to have a performance-enhancing effect on experts in the workplace, it appears crucial that the system performs at least as well as the human experts on their own. However, the individual performance of the AI should ideally surpass that of the human experts to achieve moderate or significant improvements in the performance of experts. Based on this information, it could be concluded that an ideal scenario for implementing AI occurs when the average human performance on a task is moderate, offering substantial room for improvement, and the AI introduced operates at a significantly higher performance level. However, in everyday work, the goal of introducing an AI-DSS is not merely to surpass individual human performance, but also to outperform the AI system on its own. Otherwise, from a decision-theoretic perspective, it would only be rational to delegate the task entirely to the AI (Bansal et al., 2021). Therefore, in the second part of RQ1a, we explored whether and to what extent the combined performance of humans and AI exceeds that of the AI alone. To address this part of RQ1a, we could only identify ten studies that reported on the individual performance of both the AI system and the human expert, as well as their combined performance. Unfortunately, these studies did not explicitly examine whether a significant difference exists between the individual performance of the AI and the combined performance of humans and AI. Therefore, the data did not allow for a definite answer to this sub-question, but it did enable us to make the following observations: As discussed previously, when the AI system maintained high performance while the human performance was significantly lower or moderate, the collective performance remained higher than the human performance alone but lagged behind the performance of the AI system (Kavya et al., 2021; Rajpurkar et al., 2020; Rudie et al., 2021). However, when humans and AI systems had approximately comparable individual performances, their interaction resulted in a performance that exceeded that of both the human and system (Bai et al., 2020; Kiani et al., 2020; Yang et al., 2021). This observation supports the assumption of Bansal et al. (2021) in which complementary performance is based on comparable individual performance. However, in an exploratory analysis of studies with laypersons, for a similar performance relationship between human and machine, the joint performance did not exceed both individual performances, but merely improved that of the human (see, e.g., Green and Chen, 2019a, 2019b). This can be attributed to the fact that, in the expert studies in our review, the performance of humans and AI was not only comparable, but also generally high, in contrast to the studies that included laypersons. Thus, the participants in these studies were probably better able to judge when the advice of an AI was correct or incorrect owing to their high level of expertise, as evidenced by their higher performance scores. Detailed research by Gaube et al. (2021) on how humans deal with correct and incorrect AI advice support this assumption. The study showed that less experienced and competent experts have more difficulty identifying and overriding incorrect AI advice than their more experienced or competent colleagues. However, even the latter group still tends to over-rely on AI systems, albeit to a lesser degree (Jacobs et al., 2021; Jussupow et al., 2021; Kiani et al., 2020; Lacroux and Martin-Lacroux, 2022). Thus, in relation to RQ2, it can be concluded that the accuracy of an AI-DSS significantly influences the performance of experts in work-related decision-making situations (see Subsection 3.3). Whether the automation bias (Lee and See, 2004; Skitka et al., 1999) will persist with prolonged use of the system or users will learn to evaluate the performance of the AI more accurately over time remains an open question. In addition, the question of the effects of this system property on the psychological load of the system users remains unanswered. Current research is focused on helping users to better understand when to trust and not trust AI system advice. One of the most popular approaches is to provide additional system explanations to make the inherently opaque systems more understandable to the users, and therefore, easier to judge (Lai et al., 2023). For this purpose, two types of explanations are typically developed in practice using XAI methods: decision and model explanations. Decision explanations should help users to understand individual data-related decisions more precisely, which is referred to as data explainability. Model explanations are intended to help to understand the model interdependencies. This involves the general functional relationships between the input and output variables (Kraus et al., 2021). The identified studies exclusively investigated the performance-improving effect of decision explanations on experts. They consistently showed that they are helpful for novices (or non-task experts). For senior experts, the few identified studies did not show such clear results (Gaube et al., 2023; Hwang et al., 2022; Yoon et al., 2023); hence, we cannot answer this part of RQ2, and further research is required. Specifically, it would be interesting to investigate further the assumption that experienced experts also benefit from explanations, but only in more complex cases in which their heuristics are no longer sufficient and additional explanations become necessary. It will also be worthwhile to explore how different explanatory approaches affect experts in future studies, as the number of studies on this topic remains very limited. In addition, future studies should analyse the impact of model explanations. This is because, according to qualitative research, users intend to understand the local, case-specific reasons for the model decisions and the fundamental and global properties of the model, such as its known strengths and limitations, as well as overarching design goals, that is, what it should be optimised for (Nourani et al., 2022; Riveiro and Thill, 2022). According to Riveiro and Thill (2022), a significant number of users also request explanations only when the system behaves differently from their perspective. Therefore, in the work context, investigating whether users react differently to systems that continuously display explanations compared with systems that only provide explanations on request will be interesting.

4.2 Limitations and general implications for future research

Our study had certain limitations. Most of the identified studies investigate the effectiveness of AI-DSS in the medical field (see Table 4). Consequently, the findings obtained can primarily be applied to medical cases. To test the applicability of the results to different professional contexts, the effects of AI-DSS on users in wider professional settings, such as finance, should be investigated. For a specific area in the medical field, which was not regarded in our study owing to our broad application focus, we recommend the use of specialised medical databases such as PubMed. In addition, we may have missed relevant studies owing to our search strategy. We focused on publications from 2018 onwards, as AI technologies have recently reached a new level of maturity owing to significant technical advances (Lai et al., 2023; Levy et al., 2021; Nicodeme, 2020) and are currently perceived as advanced and powerful (Almarashda et al., 2022). This perception may strongly influence human reaction towards AI-DSS (Liu et al., 2023). However, relevant research may also have been published before this period and is missing from our overview. Therefore, our findings reflect the most recent research. Third, our screening strategy focused exclusively on controlled experimental studies. This decision is based on the conviction that this methodological gold standard is best suited to identify causal relationships (Sharma et al., 2020). Compared to alternative approaches, such as a pre-post design, it allows for effective control of confounding variables. However, controlled experiments are usually conducted in simulated environments rather than in reality, as was the case in most of the included studies. Thus, users may behave differently in such environments than in natural situations. For example, in the included studies, although user performance improved significantly with AI support, users did not feel more confident in their decisions (Finck et al., 2022; Zhou et al., 2021). This finding may indicate that in simulated environments users are more willing to trust a system because they are not confronted with the consequences of the real world, even if they do not feel confident. The open question is whether they would show this behaviour in a real situation. Therefore, in future, a more controlled field research should be conducted to better understand the human-AI interactions. The results of our study provide an excellent basis for this. Fourth, the 44 included studies often had small sample sizes. This is presumably due to the difficulties in recruiting experts as study participants, owing to which some researchers did not report significance values. Consequently, interpreting the findings of these studies requires a careful approach, as we have adopted in our analysis. In the future, it will be desirable to conduct a larger number of studies with larger expert samples and consistent research designs to enable meta-analyses. In particular, experiments using a within-subject design should incorporate standardised washout periods if a between-subject design is not feasible, which would be the preferred approach. In addition, all studies should more clearly document whether participants were provided with information about the performance of the AI. It would also be highly beneficial for studies to report AI performance metrics, such as accuracy.

Furthermore, future studies should focus on more complex, multi-categorical problems rather than limiting themselves to binary decisions (e.g., diseased vs. not diseased). In this context, it would also be interesting to investigate how homogeneous user groups respond to AI systems with varying levels of accuracy. In addition, investigating whether users change their reaction to AI advice over time if they receive performance feedback during the experiment will be interesting in the future. This aspect has not been addressed in the studies considered thus far. Furthermore, we hope that future reviews will examine the effectiveness of other system design characteristics, such as cognitive forcing approaches (Buçinca et al., 2021; Jussupow et al., 2021; Langer et al., 2021) or uncertainty communication measures (He et al., 2023; Prabhudesai et al., 2023), which are also intended to promote constructive engagement with AI recommendations.

4.3 Practical implications

In addition to the need for further research, specific recommendations for using AI-DSS in professional practice can be derived from the summarised and discussed study results. First, organisations planning to use AI-DSS should carefully analyse and understand the potential implementation context of the tool. Specifically, clarifying whether the tasks supported by the technical system are perceived as sufficiently demanding and complex by employees is essential. This is because technical assistance can only offer tangible added value if there is a mismatch between existing mental requirements for the decision-making process and the existing resources of the decision-makers (Langer et al., 2021). However, it is noteworthy that even small performance improvements of several percentage points can be considered significant, especially in critical areas such as medicine. Moreover, our results also suggest that, in areas where performance is already high, collaboration between humans and AI is particularly promising, as they appear to complement one another synergistically and outperform their individual capabilities – provided that the AI performs at a comparable level to that of a human. However, if the system is significantly below human performance, the performance may deteriorate owing to the technical support. This is because, as the review results show, both experts, and particularly novices and non-experts, tend to initially overtrust AI-DSS. Organisations should, therefore, also take additional measures when designing their systems to help develop a more appropriate level of trust in AI systems. System explanations have already proven to be an effective mean of achieving this for the target group of novices and non-experts. However, research in this area is still in its infancy; thus, we strongly recommend that organisations conduct internal user tests on the effect of various explanatory approaches or other transparency measures, such as uncertainty communication or cognitive forcing strategies. In addition, organisations should also ensure that general usability design criteria, such as simplicity, are adhered to when designing interfaces (Lee et al., 2007). Otherwise, there is a risk seen that users will be cognitively overwhelmed by extensive explanations, and the intended relief effects of AI support can be cancelled out (Pushparaj et al., 2023). In summary, systems should be developed, implemented, and evaluated according to the human-centred design approach (ISO International Organization for Standardization, 2019). This ensures that the specific needs of future users are at the centre of attention and that the human-AI interaction is successful.

5 Conclusions

The extensive literature review shows that current research on AI-DSS mainly investigates the effects of these technologies on expert performance, often with a focus on the medical domain. The combined results of the studies suggest that human-AI interaction in work-related decision situations can significantly improve expert performance, compared with the average performance of both: their own and system alone. Whether this happens in individual cases depends largely on the level of individual human and system performance and their interaction. First, the added value of human-AI interaction is limited if, for example, the performance of the individual expert is already significantly high. Second, the results show that experts are dependent on the performance level of the AI system, as they also have difficulties recognising incorrect AI advice. Notably, this applies more strongly to inexperienced and less capable experts. The results suggest that superior human-AI performance, surpassing that of a single entity, requires relatively high and complementary individual performance from both. However, given the limited number of studies reporting all three accuracy values and the small sample sizes, further research is needed to draw general conclusions about the required levels of system and human accuracy and their ratio to each other. This also applies to the role of system explainability. Previous research has shown that explanations of individual system decisions help novices make appropriate decisions. However, the influence on experienced experts is still unclear and the effect of global explanations remains unexplored. The summarised study findings originate mainly from simulated experiments, underlining the need to verify external validity through future research in real work contexts.

Acknowledgements

The research presented in this paper was carried out within the research project ‘AIXPERIMENTATIONLAB’ (Project number EXP.01.00016.20). The authors gratefully acknowledge the support of the German Federal Ministry of Labor and Social Affairs (BMAS) and Initiative New Quality of Work (INQA).

Data availability statement

The authors are pleased to offer access to the data collected during this systematic review. Please contact us via email, if interested. We will then gladly invite you to join the relevant project in Rayyan, where the literature review was conducted.

References

Ajzen, I., Fishbein, M., Lohmann, S. and Albarracín, D. (2018) ‘The influence of attitudes on behavior’, in Albarracin, D. and Johnson, B.T. (Eds.): *The Handbook of Attitudes, Volume 1: Basic Principles*, Routledge, New York.

Almarashda, H.A.H.A., Baba, I.B., Ramli, A.A., Memon, A.H. (2022) ‘User expectation and benefits of implementing artificial intelligence in the UAE energy sector’, *Journal of Applied Engineering Sciences*, Vol. 12, pp.1–10, https://doi.org/10.2478/jaes-2022-0001.

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. and Herrera, F. (2020) 'Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI', *Information Fusion*, Vol. 58, pp.82–115, https://doi.org/10.1016/j.inffus.2019.12.012.

Bai, H.X., Wang, R., Xiong, Z., Hsieh, B., Chang, K., Halsey, K., Tran, T.M.L., Choi, J.W., Wang, D-C., Shi, L-B., Mei, J., Jiang, X-L., Pan, I., Zeng, Q-H., Hu, P-F., Li, Y-H., Fu, F-X., Huang, R.Y., Sebro, R., Yu, Q-Z., Atalay, M.K. and Liao, W-H. (2020) 'Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT', *Radiology*, Vol. 296, pp.156–166, https://doi.org/10.1148/radiol.2020201491.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M.T. and 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, CHI '21*, Association for Computing Machinery, New York, USA, pp.1–16, https://doi.org/10.1145/3411764.3445717.

Bayer, S., Gimpel, H. and Markgraf, M. (2022) 'The role of domain expertise in trusting and following explainable AI decision support systems', *Journal of Decision Systems*, Vol. 32, pp.110–138, https://doi.org/10.1080/12460125.2021.1958505.

Beijaard, D., Verloop, N. and Vermunt, J.D. (2000) 'Teachers' perceptions of professional identity: an exploratory study from a personal knowledge perspective', *Teaching and Teaching Education*, Vol. 16, No. 7, pp.749–764, https://doi.org/10.1016/S0742-051X(00)00023-8.

Bhatt, U., Weller, A. and Moura, J.M.F. (2020) 'Evaluating and aggregating feature-based model explanations', in *IJCAI '20: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, Association for Computing Machinery*, New York, USA, pp.3016–3022, https://doi.org/10.48550/arXiv.2005.00631.

Buçinca, Z., Lin, P., Gajos, K.Z. and Glassman, E.L. (2020) 'Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems', in *IUI '20: Proceedings of the 25th International Conference on Intelligent User Interfaces*, Association for Computing Machinery, New York, USA, pp.454–464, https://doi.org/10.1145/3377325.3377498.

Buçinca, Z., Malaya, M.B. and Gajos, K.Z. (2021) 'To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making', in *Proceedings of the ACM on Human-Computer Interaction*, Association for Computing Machinery, New York, USA, Vol. 5, No. 188, pp.1–21, https://doi.org/10.1145/3449287.

Cai, C.J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., Wattenberg, M., Viegas, F., Corrado, G.S., Stumpe, M.C. and Terry, M. (2019) 'Human-centered tools for coping with imperfect algorithms during medical decision-making', in *CHI 2019: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, New York, USA, pp.1–14, https://doi.org/10.1145/3290605.3300234.

Calisto, F.M., Santiago, C., Nunes, N. and Nascimento, J.C. (2022) 'BreastScreening-AI: evaluating medical intelligent agents for human-AI interactions', *Artificial Intelligence in Medicine*, Vol. 127, p.102285, https://doi.org/10.1016/j.artmed.2022.102285.

Carton, S., Mei, Q. and Resnick, P. (2020) 'Feature-based explanations don't help people detect misclassifications of online toxicity', in *Proceedings of the International AAAI Conference on Web and Social Media*, Association for the Advancement of Artificial Intelligence, Palo Alto, USA, Vol. 14, pp.95–106, https://doi.org/10.1609/icwsm.v14i1.7282.

Chen, V., Liao, Q.V., Wortman Vaughan, J. and Bansal, G. (2023) 'Understanding the role of human intuition on reliance in human-AI decision-making with explanations', in *Proceedings of the ACM on Human-Computer Interaction*, Association for Computing Machinery, New York, USA, Vol. 7, pp.1–32, https://doi.org/10.1145/3610219.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed., L. Erlbaum Associates, Hillsdale.

Didimo, W., Giamminonni, L., Liotta, G., Montecchiani, F. and Pagliuca, D. (2018) ‘A visual analytics system to support tax evasion discovery’, *Decision Support Systems*, Vol. 110, pp.71–83, https://doi.org/10.1016/j.dss.2018.03.008.

Dikmen, M. and Burns, C. (2022) ‘The effects of domain knowledge on trust in explainable AI and task performance: a case of peer-to-peer lending’, *International Journal of Human-Computer Studies*, Vol. 162, p.102792, https://doi.org/10.1016/j.ijhcs.2022.102792.

Dorr, F., Chaves, H., Serra, M.M., Ramirez, A., Costa, M.E., Seia, J. and Barmaimon, G. (2020) ‘COVID-19 pneumonia accurately detected on chest radiographs with artificial intelligence’, *Intelligence-Based Medicine*, p.100014, https://doi.org/10.1016/j.ibmed.2020.100014.

Duchevet, A., Imbert, J-P., La Hogue, T.D., Ferreira, A., Moens, L., Colomer, A., Cantero, J., Bejarano, C. and Vázquez, A.L.R. (2022) ‘HARVIS: a digital assistant based on cognitive computing for non-stabilized approaches in single pilot operations’, *Transportation Research Procedia*, Vol. 66, pp.253–261, https://doi.org/10.1016/j.trpro.2022.12.025.

Finck, T., Moosbauer, J., Probst, M., Schlaeger, S., Schuberth, M., Schinz, D., Yiğitsoy, M., Byas, S., Zimmer, C., Pfister, F. and Wiestler, B. (2022) ‘Faster and better: how anomaly detection can accelerate and improve reporting of head computed tomography’, *Diagnostics*, Vol. 12, p.452, https://doi.org/10.3390/diagnostics12020452.

Gartner (2023) *Gartner Poll Finds 45% of Executives Say ChatGPT Has Prompted an Increase in AI Investment* [online] https://www.gartner.com/en/newsroom/press-releases/2023-05-03-gartner-poll-finds-45-percent-of-executives-say-chatgpt-has-prompted-an-increase-in-ai-investment (accessed 28 February 2024).

Gaube, S., Suresh, H., Raue, M., Lermer, E., Koch, T.K., Hudecek, M.F.C., Ackery, A.D., Grover, S.C., Coughlin, J.F., Frey, D., Kitamura, F.C., Ghassemi, M. and Colak, E. (2023) ‘Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays’, *Scientific Reports*, Vol. 13, No. 1, p.1383, https://doi.org/10.1038/s41598-023-28633-w.

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S.J., Lermer, E., Coughlin, J.F., Guttag, J.V., Colak, E. and Ghassemi, M. (2021) ‘Do as AI say: susceptibility in deployment of clinical decision-aids’, *NPJ Digital Medicine*, Vol. 4, No. 1, p.31, https://doi.org/10.1038/s41746-021-00385-9.

Green, B. and Chen, Y. (2019a) ‘Disparate interactions: an algorithm-in-the-loop analysis of fairness in risk assessments’, in *FAT* ‘19: Proceedings of the Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, New York, USA, https://doi.org/10.1145/3287560.3287563.

Green, B. and Chen, Y. (2019b) ‘The principles and limits of algorithm-in-the-loop decision making’, in *Proceedings of the ACM on Human-Computer Interaction*, Association for Computing Machinery, New York, USA, Vol. 3, No. 50, pp.1–24, https://doi.org/10.1145/3359152.

He, G., Buijsman, S. and Gadiraju, U. (2023) ‘How stated accuracy of an AI system and analogies to explain accuracy affect human reliance on the system’, in *Proceedings of the ACM on Human-Computer Interaction*, Association for Computing Machinery, New York, USA, Vol. 7, No. 276, pp.1–29, https://doi.org/10.1145/3610067.

Hellebrandt, T., Huebser, L., Adam, T., Heine, I. and Schmitt, R.H. (2021) ‘Augmented intelligence – Mensch trifft Künstliche Intelligenz: Intelligentes Zusammenwirken von Mensch und KI für bessere Entscheidungen und Handlungen in der Produktion’, *Zeitschrift für wirtschaftlichen Fabrikbetrieb*, Vol. 116, pp.433–437, https://doi.org/10.1515/zwf-2021-0104.

Henkel, M., Horn, T., Leboutte, F., Trotsenko, P., Dugas, S.G., Sutter, S.U., Ficht, G., Engesser, C., Matthias, M., Stalder, A., Ebbing, J., Cornford, P., Seifert, H., Stieltjes, B. and Wetterauer, C. (2022) ‘Initial experience with AI pathway companion: evaluation of dashboard-enhanced clinical decision making in prostate cancer screening’, *PLOS One*, Vol. 17, p.e0271183, https://doi.org/10.1371/journal.pone.0271183.

Hornung, O. and Smolnik, S. (2022) 'AI invading the workplace: negative emotions towards the organizational use of personal virtual assistants', *Electron Markets*, Vol. 32, pp.123–138, https://doi.org/10.1007/s12525-021-00493-0.

Hwang, J., Lee, T., Lee, H. and Byun, S. (2022) 'A clinical decision support system for sleep staging tasks with explanations from artificial intelligence: user-centered design and evaluation study', *Journal of Medical Internet Research*, Vol. 24, No. 1, p.e28659, https://doi.org/10.1177/016555159902500305.

ISO International Organization for Standardization (2019) *ISO 9241-210:2010(en): Ergonomics of Human-System Interaction: Human-Centred Design for Interactive Systems* [online] https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-1:v1:en (accessed 18 March 2024).

Jacobs, M., Pradier, M.F., McCoy, T.H., Perlis, R.H., Doshi-Velez, F. and Gajos, K.Z. (2021) 'How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection', *Translational Psychiatry*, Vol. 11, pp.1–9, https://doi.org/10.1038/s41398-021-01224-x.

Janiesch, C., Zschech, P. and Heinrich, K. (2021) 'Machine learning and deep learning', *Electron Markets*, Vol. 31, pp.685–695, https://doi.org/10.1007/s12525-021-00475-2.

Jarrahi, M.H. (2018) 'Artificial intelligence and the future of work: human-AI symbiosis in organizational decision making', *Business Horizons*, Vol. 61, pp.577–586, https://doi.org/10.1016/j.bushor.2018.03.007.

Jussupow, E., Spohrer, K., Heinzl, A. and Gawlitza, J. (2021) 'Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence', *Information Systems Research*, Vol. 32, pp.713–735, https://doi.org/10.1287/isre.2020.0980.

Kavya, R., Christopher, J., Panda, S. and Lazarus, Y.B. (2021) 'Machine learning and XAI approaches for allergy diagnosis', *Biomedical Signal Processing and Control*, Vol. 69, p.102681, https://doi.org/10.1016/j.bspc.2021.102681.

Kiani, A., Uyumazturk, B., Rajpurkar, P., Wang, A., Gao, R., Jones, E., Yu, Y., Langlotz, C.P., Ball, R.L., Montine, T.J., Martin, B.A., Berry, G.J., Ozawa, M.G., Hazard, F.K., Brown, R.A., Chen, S.B., Wood, M., Allard, L.S., Ylagan, L., Ng, A.Y. and Shen, J. (2020) 'Impact of a deep learning assistant on the histopathologic classification of liver cancer', *NPJ Digital Medicine*, Vol. 3, pp.1–8, https://doi.org/10.1038/s41746-020-0232-8.

Kiefer, G-L., Safi, T., Nadig, M., Sharma, M., Sakha, M.M., Ndiaye, A., Deru, M., Daas, L., Schulz, K., Schwarz, M., Seitz, B. and Alexandersson, J. (2022) 'An AI-based decision support system for quality control applied to the use case donor cornea', in *Artificial Intelligence in HCI: 3rd International Conference, AI-HCI 2022, Held as Part of the 24th HCI International Conference, HCII 2022*, Virtual Event, Springer, Cham, Switzerland, 26 June–1 July, pp.257–274, https://doi.org/10.1007/978-3-031-05643-7_17.

Kim, H.Y., Lim, D.Y. and Song, S. (2023) 'Understanding satisfaction factors of personalized body-weight exercises', in *CSCW '23 Companion: Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, Association for Computing Machinery, New York, USA, pp.101–104, https://doi.org/10.1145/3584931.3606963.

Kim, J.H., Kim, J.Y., Kim, G.H., Kang, D., Kim, I.J., Seo, J., Andrews, J.R. and Park, C.M. (2020) 'Clinical validation of a deep learning algorithm for detection of pneumonia on chest radiographs in emergency department patients with acute febrile respiratory illness', *Journal of Clinical Medicine*, Vol. 9, p.1981, https://doi.org/10.3390/jcm9061981.

Kindler, C., Elfwing, S., Öhrvik, J. and Nikberg, M. (2023) 'A deep neural network-based decision support tool for the detection of lymph node metastases in colorectal cancer specimens', *Modern Pathology*, Vol. 36, No. 2, p100015, https://doi.org/10.1016/j.modpat.2022.100015.

Koo, Y.H., Shin, K.E., Park, J.S., Lee, J.W., Byun, S. and Lee, H. (2021) 'Extravalidation and reproducibility results of a commercial deep learning-based automatic detection algorithm for pulmonary nodules on chest radiographs at tertiary hospital', *Journal of Medical Imaging and Radiation Oncology*, Vol. 65, pp.15–22, https://doi.org/10.1111/1754-9485.13105.

Kozuka, T., Matsukubo, Y., Kadoba, T., Oda, T., Suzuki, A., Hyodo, T., Im, S., Kaida, H., Yagyu, Y., Tsurusaki, M., Matsuki, M. and Ishii, K. (2020) 'Efficiency of a computer-aided diagnosis (CAD) system with deep learning in detection of pulmonary nodules on 1-mm-thick images of computed tomography', *Japanese Journal of Radiology*, Vol. 38, pp.1052–1061, https://doi.org/10.1007/s11604-020-01009-0.

Kraus, T., Ganschow, L., Eisenträger, M. and Wischmann, S. (2021) *Erklärbare KI – Anforderungen, Anwendungsfälle und Lösungen* [online] https://scholar.google.de/citations?user=vw_91keaaaaj&hl=de&oi=sra (accessed 2 April 2024).

Lacroux, A. and Martin-Lacroux, C. (2022) 'Should i trust the artificial intelligence to recruit? Recruiters' perceptions and behavior when faced with algorithm-based recommendation systems during resume screening', *Frontiers in Psychology*, Vol. 13, p.895997, https://doi.org/10.3389/fpsyg.2022.895997.

Lai, V. and Tan, C. (2019) 'On human predictions with explanations and predictions of machine learning models: a case study on deception detection', in *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, New York, USA, pp.29–38, https://doi.org/10.1145/3287560.3287590.

Lai, V., Chen, C., Smith-Renner, A., Liao, Q.V. and Tan, C. (2023) 'Towards a science of human-AI decision making: an overview of design space in empirical human-subject studies', in *FACCT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, New York, USA, pp.1369–1385, https://doi.org/10.1145/3593013.3594087.

Lai, V., Liu, H. and Tan, C. (2020) ''Why is 'Chicago' deceptive?' Towards building model-driven tutorials for humans', in *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, New York, USA, pp.1–13, https://doi.org/10.1145/3313831.3376873.

Langer, M., König, C.J. and Busch, V. (2021) 'Changing the means of managerial work: effects of automated decision support systems on personnel selection tasks', *Journal of Business and Psychology*, Vol. 36, pp.751–769, https://doi.org/10.1007/s10869-020-09711-6.

Laursen, M.S., Pedersen, J.S., Hansen, R.S., Savarimuthu, T.R., Lynggaard, R.B. and Vinholt, P.J. (2023) 'Doctors identify hemorrhage better during chart review when assisted by artificial intelligence', *Applied clinical informatics*, Vol. 14, pp.743–751, https://doi.org/10.1055/a-2121-8380

Lee, D., Moon, J. and Kim, Y. (2007) 'The effect of simplicity and perceived control on perceived ease of use', in *AMCIS 2007: Americas Conference on Information Systems*, Association for Information Systems, Atlanta, USA, Vol. 71, pp.1–16 [online] https://aisel.aisnet.org/amcis2007/71 (accessed 22 April 2024).

Lee, J.D. and See, K.A. (2004) 'Trust in automation: designing for appropriate reliance', *Human Factors*, Vol. 46, No. 1, pp.50–80, https://doi.org/10.1518/hfes.46.1.50_30392.

Lee, M.H. and Chew, C.J. (2023) 'Understanding the effect of counterfactual explanations on trust and reliance on AI for human-AI collaborative clinical decision making', in *Proceedings of the ACM on Human-Computer Interaction. Association for Computing Machinery*, New York, USA, Vol. 7, No. 369, pp.1–22, https://doi.org/10.1145/3610218.

Lee, M.H., Siewiorek, D.P., Smailagic, A., Bernardino, A. and Bermúdez i Badia, S. (2020) 'Co-design and evaluation of an intelligent decision support system for stroke rehabilitation assessment', in *Proceedings of the ACM on Human-Computer Interaction*, Association for Computing Machinery, New York, USA, Vol. 4, No. 5, pp.1–27, https://doi.org/10.1145/3415227.

Levy, A., Agrawal, M., Satyanarayan, A. and Sontag, D. (2021) 'Assessing the impact of automated suggestions on decision making: domain experts mediate model errors but take less initiative', in *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, New York, USA, pp.1–13, https://doi.org/10.1145/3411764.3445522.

Li, D., Pehrson, L.M., Lauridsen, C.A., Tøttrup, L., Fraccaro, M., Elliott, D., Zając, H.D., Darkner, S., Carlsen, J.F. and Nielsen, M.B. (2021) 'The added effect of artificial intelligence on physicians' performance in detecting thoracic pathologies on CT and chest X-ray: a systematic review', *Diagnostics*, Vol. 11, p.2206, https://doi.org/10.3390/diagnostics11122206.

Lin, Z.J., Jung, J., Goel, S. and Skeem, J. (2020) 'The limits of human predictions of recidivism', *Science Advances*, Vol. 6, No.7, p.eaaz0652, https://doi.org/10.1126/sciadv.aaz0652.

Liu, B. (2021) 'In AI we trust? Effects of agency locus and transparency on uncertainty reduction in human-AI interaction', *Journal of Computer-Mediated Communication*, Vol. 26, pp.384–402, https://doi.org/10.1093/jcmc/zmab013.

Liu, H., Lai, V. and Tan, C. (2021) 'Understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making', in *Proceedings of the ACM on Human-Computer Interaction*, Association for Computing Machinery, New York, USA, Vol. 5, No. 408, pp.1–45, https://doi.org/10.1145/3479552.

Liu, K., Li, Q., Ma, J., Zhou, Z., Sun, M., Deng, Y., Tu, W., Wang, Y., Fan, L., Xia, C., Xiao, Y., Zhang, R. and Liu, S. (2019) 'Evaluating a fully automated pulmonary nodule detection approach and its impact on radiologist performance', *Radiology: Artificial intelligence*, Vol. 1, No. 3, p.e180084, https://doi.org/10.1148/ryai.2019180084.

Liu, M., Ke, W. and Xu, D.J. (2023) 'Will humans be free-riders? The effects of expectations for AI on human-AI team performance', in *PACIS 2023 Proceedings: Pacific Asia Conference on Information Systems*, Association for Information Systems, Nanchang, China, Vol. 20 [online] https://aisel.aisnet.org/pacis2023/20 (accessed 4 January 2024).

Marois, R. and Ivanoff, J. (2005) 'Capacity limits of information processing in the brain', *Trends in Cognitive Sciences*, Vol. 9, No. 6, pp.296–305, https://doi.org/10.1016/j.tics.2005.04.010.

Martini, K., Blüthgen, C., Eberhard, M., Schönenberger, A.L.N., Martini, I., Huber, F.A., Barth, B.K., Euler, A. and Frauenfelder, T. (2021) 'Impact of vessel suppressed-CT on diagnostic accuracy in detection of pulmonary metastasis and reading time', *Academic Radiology*, Vol. 28, No. 7, pp.988–994, https://doi.org/10.1016/j.acra.2020.01.014.

Mercado, J.E., Rupp, M.A., Chen, J.Y.C., Barnes, M.J., Barber, D. and Procci, K. (2016) 'Intelligent agent transparency in human-agent teaming for multi-UxV management', *Human Factors: The Journal of the Human Factors and Ergonomics Society*, Vol. 58, No. 3, pp.401–415, https://doi.org/10.1177/0018720815621206.

Murphy, K.P. (2012) 'Machine learning, a probabilistic perspective', *Adaptive Computation and Machine Learning Series*, MIT Press, London.

Naikar, N., Brady, A., Moy, G. and Kwok, H-W. (2023) 'Designing human-AI systems for complex environments: ideas from distributed, joint, and self-organising perspectives of sociotechnical systems and cognitive work analysis', *Ergonomics*, Vol. 66, No. 11, pp.1669–1694, https://doi.org/10.1080/00140139.2023.2281898.

Nam, J.G., Kim, M., Park, J., Hwang, E.J., Lee, J.H., Hong, J.H., Goo, J.M. and Park, C.M. (2021) 'Development and validation of a deep learning algorithm detecting 10 common abnormalities on chest radiographs', *European Respiratory Journal*, Vol. 57, https://doi.org/10.1183/13993003.03061-2020.

Nicodeme, C. (2020) 'Build confidence and acceptance of AI-based decision support systems – explainable and liable AI', in *2020 13th International Conference on Human System Interaction (HSI)*, Institute of Electrical and Electronics Engineers (IEEE), Tokyo, Japan, pp.20–23, https://doi.org/10.1109/hsi49210.2020.9142668.

Nourani, M., Roy, C., Block, J.E., Honeycutt, D.R., Rahman, T., Ragan, E.D. and Gogate, V. (2022) 'On the importance of user backgrounds and impressions: lessons learned from interactive AI applications', *ACM Transactions on Interactive Intelligent Systems*, Vol. 12, No. 28, pp.1–29, https://doi.org/10.1145/3531066.

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D. and Moher, D. (2021) 'The PRISMA 2020 statement: an updated guideline for reporting systematic reviews', *BMJ*, Vol. 372, No. 71, https://doi.org/10.1136/bmj.n71.

Panigutti, C., Beretta, A., Giannotti, F. and Pedreschi, D. (2022) 'Understanding the impact of explanations on advice-taking: a user study for AI-based clinical decision support systems', in *CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, New York, USA, No. 568, https://doi.org/10.1145/3491102.3502104.

Popescu, C., Golden, G., Benrimoh, D., Tanguay-Sela, M., Slowey, D., Lundrigan, E. and Turecki, G. (2021) 'Evaluating the clinical feasibility of an artificial intelligence – powered, web-based clinical decision support system for the treatment of depression in adults: longitudinal feasibility study', *JMIR Formative Research*, Vol. 5, p.e31862, https://doi.org/10.2196/31862.

Poursabzi-Sangdeh, F., Goldstein, D.G., Hofman, J.M., Wortman Vaughan, J.W. and Wallach, H. (2021) 'Manipulating and measuring model interpretability', in *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, New York, USA, No. 237, pp.1–52, https://doi.org/10.1145/3411764.3445315.

Prabhudesai, S., Yang, L., Asthana, S., Huan, X., Liao, Q.V. and Banovic, N. (2023) 'Understanding uncertainty: how lay decision-makers perceive and interpret uncertainty in human-AI decision making', in *IUI '23: Proceedings of the 28th International Conference on Intelligent User Interfaces*, Association for Computing Machinery, New York, USA, pp.379–396, https://doi.org/10.1145/3581641.3584033.

Pushparaj, K., Reddy, P., Vu-Tran, D., Izzetoglu, K. and Alam, S. (2023) 'A multi-modal approach to measuring the effect of XAI on air traffic controller trust during off-nominal runway exits', in *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Institute of Electrical and Electronics Engineers (IEEE), Honolulu, USA, pp.4813–4819, https://doi.org/10.1109/SMC53992.2023.10394443.

Rajpurkar, P., O'Connell, C., Schechter, A., Asnani, N., Li, J., Kiani, A., Ball, R.L., Mendelson, M., Maartens, G., van Hoving, D.J., Griesel, R., Ng, A.Y., Boyles, T.H. and Lungren, M.P. (2020) 'CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV', *NPJ Digital Medicine*, Vol. 3, No. 115, pp.1–8, https://doi.org/10.1038/s41746-020-00322-2.

Riveiro, M. and Thill, S. (2022) 'The challenges of providing explanations of AI systems when they do not behave like users expect', in *UMAP '22: Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, Association for Computing Machinery, New York, USA, pp.110–120, https://doi.org/10.1145/3503252.3531306.

Rodriguez-Ruiz, A., Jan-Jurre, M., Karssemeijer, N., Sechopoulos, I. and Mann, R.M. (2018) 'Can radiologists improve their breast cancer detection in mammography when using a deep learning based computer system as decision support?', in *Proceedings of SPIE Volume 10718: 14th International Workshop on Breast Imaging (IWBI 2018)*, SPIE, pp.7–16, https://doi.org/10.1117/12.2317937.

Roller, R., Mayrdorfer, M., Duettmann, W., Naik, M.G., Schmidt, D., Halleck, F., Hummel, P., Burchardt, A., Möller, S., Dabrock, P., Osmanodja, B. and Budde, K. (2022) 'Evaluation of a clinical decision support system for detection of patients at risk after kidney transplantation', *Frontiers in Public Health*, Vol. 10, p.979448, https://doi.org/10.3389/fpubh.2022.979448.

Rudie, J.D., Duda, J., Duong, M.T., Chen, P-H., Xie, L., Kurtz, R., Ware, J.B., Choi, J., Mattay, R.R., Botzolakis, E.J., Gee, J.C., Bryan, R.N., Cook, T.S., Mohan, S., Nasrallah, I.M. and Rauschecker, A.M. (2021) 'Brain MRI deep learning and Bayesian inference system augments radiology resident performance', *Journal of Digital Imaging*, Vol. 34, pp.1049–1058, https://doi.org/10.1007/s10278-021-00470-1.

Shah, C., Davtyan, K., Nasrallah, I., Bryan, R.N. and Mohan, S. (2023) 'Artificial intelligence-powered clinical decision support and simulation platform for radiology trainee education', *Journal of Digital Imaging*, Vol. 36, pp.11–16, https://doi.org/10.1007/s10278-022-00713-9.

Sharma, N., Srivastav, A.K. and Samuel, A.J. (2020) 'Randomized clinical trial: gold standard of experimental designs – importance, advantages, disadvantages and prejudices', *Revista Pesquisa em Fisioterapia*, Vol. 10, pp.512–519, https://doi.org/10.17267/2238-2704rpf.v10i3.3039.

Singh, P., Bhardwaj, P., Sharma, S.K. and Agrawal, A.K. (2022) 'Psychological stress and job satisfaction in middle management executives: a test of job demand control support model', *International Journal of Human Factors and Ergonomics*, Vol. 9, No. 4, pp.372–388, https://doi.org/10.1504/IJHFE.2022.127447.

Singh, R., Kalra, M.K., Homayounieh, F., Nitiwarangkul, C., McDermott, S., Little, B.P., Lennes, I.T., Shepard, J-A.O. and Digumarthy, S.R. (2021) 'Artificial intelligence-based vessel suppression for detection of sub-solid nodules in lung cancer screening computed tomography', *Quantitative Imaging in Medicine and Surgery*, Vol. 11, pp.1134–1143, https://doi.org/10.21037/qims-20-630.

Sivaraman, V., Bukowski, L.A., Levin, J., Kahn, J.M. and Perer, A. (2023) 'Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care', in *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, New York, USA, pp.1–18, https://doi.org/10.1145/3544548.3581075.

Skitka, L.J., Mosier, K.L. and Burdick, M. (1999) 'Does automation bias decision-making?', *International Journal of Human-Computer Studies*, Vol. 51, No. 5, pp.991–1006, https://doi.org/10.1006/ijhc.1999.0252.

Spector, J.M. and Ma, S. (2019) 'Inquiry and critical thinking skills for the next generation: from artificial intelligence back to human intelligence', *Smart Learning Environments*, Vol. 6, pp.1–11, https://doi.org/10.1186/s40561-019-0088-z.

Stowers, K., Kasdaglis, N., Rupp, M.A., Newton, O.B., Chen, J.Y.C. and Barnes, M.J. (2020) 'The IMPACT of agent transparency on human performance', in *IEEE Transactions on Human-Machine Systems*, Institute of Electrical and Electronics Engineers (IEEE), Vol. 50, pp.245–253, https://doi.org/10.1109/thms.2020.2978041.

Sung, J., Park, S., Lee, S.M., Bae, W., Park, B., Jung, E., Seo, J.B. and Jung, K.-H. (2021) 'Added value of deep learning-based detection system for multiple major findings on chest radiographs: a randomized crossover study', *Radiology*, Vol. 299, pp.450–459, https://doi.org/10.1148/radiol.2021202818.

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M.S. and Krishna, R. (2023) 'Explanations can reduce overreliance on AI systems during decision-making', in *Proceedings of the ACM on Human-Computer Interaction*, Association for Computing Machinery, New York, USA, Vol. 7, No. 129, pp.1–38, https://doi.org/10.1145/3579605.

Vasconcelos, M., Cardonha, C. and Gonçalves, B. (2018) 'Modeling epistemological principles for bias mitigation in AI systems: an illustration in hiring decisions', in *AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, Association for Computing Machinery, New York, USA, pp.323–329, https://doi.org/10.1145/3278721.3278751.

von de Merwe, K., Mallam, S. and Nazir, S. (2022) 'Agent transparency, situation awareness, mental workload, and operator performance: a systematic literature review', in *Human Factors: The Journal of the Human Factors and Ergonomics Society, Human Factors and Ergonomics Society*, Vol. 66, https://doi.org/10.1177/00187208221077804.

Walczok, M. and Bipp, T. (2023) 'Investigating the effect of intelligent assistance systems on motivational work characteristics in assembly', *Journal of Intelligent Manufacturing*, pp.1–14, https://doi.org/10.1007/s10845-023-02086-4.

Wang, X. and Yin, M. (2021) 'Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making', in *IUI '21: Proceedings of the 26th International Conference on Intelligent User Interfaces*, Association for Computing Machinery, New York, USA, pp.318–328, https://doi.org/10.1145/3397481.3450650.

Weerts, H.J.P., van Ipenburg, W. and Pechenizkiy, M. (2019) *A Human-Grounded Evaluation of SHAP for Alert Processing*, https://doi.org/10.48550/arXiv.1907.03324.

Wilkens, U. (2020) 'Artificial intelligence in the workplace – a double-edged sword', *The International Journal of Information and Learning Technology,* Vol. 37, No. 5, pp.253–265, https://doi.org/10.1108/IJILT-02-2020-0022.

Wohlin, C. (2014) 'Guidelines for snowballing in systematic literature studies and a replication in software engineering', in *EASE '14: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, Association for Computing Machinery, New York, USA, No. 38, pp.1–10, https://doi.org/10.1145/2601248.2601268.

Yang, Y., Lure, F.Y.M., Miao, H., Zhang, Z., Jaeger, S., Liu, J. and Guo, L. (2021) 'Using artificial intelligence to assist radiologists in distinguishing COVID-19 from other pulmonary infections', *Journal of X-Ray Science and Technology*, Vol. 29, pp.1–17, https://doi.org/10.3233/XST-200735.

Yao, X., Rushlow, D.R., Inselman, J.W., McCoy, R.G., Thacher, T.D., Behnken, E.M., Bernard, M.E., Rosas, S.L., Akfaly, A., Misra, A., Molling, P.E., Krien, J.S., Foss, R.M., Barry, B.A., Siontis, K.C., Kapa, S., Pellikka, P.A., Lopez-Jimenez, F., Attia, Z.I., Shah, N.D., Friedman, P.A. and Noseworthy, P.A. (2021) 'Artificial intelligence–enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial', *Nature Medicine*, Vol. 27, pp.815–819, https://doi.org/10.1038/s41591-021-01335-4.

Yoon, J., Han, J., Ko, J., Choi, S., Park, J.I., Hwang, J.S., Han, J.M. and Hwang, D.D-J. (2023) 'Developing and evaluating an AI-based computer-aided diagnosis system for retinal disease: diagnostic study for central serous chorioretinopathy', *Journal of Medical Internet Research*, Vol. 25, p.e48142, https://doi.org/10.2196/48142.

Zhang, D., Liu, X., Shao, M., Sun, Y., Lian, Q. and Zhang, H. (2023) 'The value of artificial intelligence and imaging diagnosis in the fight against COVID-19', *Personal and Ubiquitous Computing*, Vol. 27, No. 3, pp.783–792, https://doi.org/10.1007/s00779-021-01522-7.

Zhang, Y., Liao, Q.V. and Bellamy, R.K.E. (2020) 'Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making', in *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, New York, USA, pp.295–305, https://doi.org/10.1145/3351095.3372852.

Zhou, G., Aggarwal, V., Yin, M. and Yu, D. (2021) 'Video-based AI decision support system for lifting risk assessment', in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Institute of Electrical and Electronics Engineers (IEEE), Melbourne, Australia, https://doi.org/10.1109/smc52423.2021.9659025.

Zhou, L., Rudin, C., Gombolay, M., Spohrer, J., Zhou, M. and Paul, S. (2023) 'From artificial intelligence (AI) to intelligence augmentation (IA): design principles, potential risks, and emerging issues', *AIS Transactions on Human-Computer Interaction*, Vol. 15, pp.111–135, https://doi.org/10.17705/1thci.00085.