



Human–AI Interface Design for Trust Calibration and Cognitive Workload Management in High Stakes Decision Making Contexts: A Scoping Review

Fisayo Fakinlede ^{a*}, Daniel Kofi Yeboah ^b
and Grace Oluwaseun Ikudehinbu ^c

^a *Information Systems and Business Analytics, College of Business, Iowa State University, Ames, Iowa, United States.*

^b *College of Professional Studies, Northeastern University, Portland, Maine, United States.*

^c *Southern Illinois University-Edwardsville School of Business, Illinois, United States.*

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: <https://doi.org/10.9734/jsrr/2026/v32i64237>

Open Peer Review History:

This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: <https://pr.sdiarticle5.com/review-history/156846>

Systematic Review Article

Received: 26/02/2026
Published: 03/06/2026

Abstract

High-stakes decision environments continue to face cognitive demands, accountability pressures, and uncertainty, even as artificial intelligence is increasingly embedded in decision support across critical infrastructure, safety-sensitive domains, healthcare, and aviation. This scoping review examines how the design of the human–AI interface shapes cognitive workload, safe reliance, and trust calibration in such settings. A PCC-framed question and a PRISMA-ScR-guided process were used to identify studies published between 2015 and 2025 in Scopus, PubMed/MEDLINE, Web of Science, ScienceDirect, IEEE Xplore, and

*Corresponding author: E-mail: fakinledefisayo52@gmail.com;

the ACM Digital Library. These were screened and charted using an extraction template. Seventeen studies were chosen, covering clinical decision support, sepsis management, medical imaging, power-grid congestion management, telehealth diagnosis, air traffic control, medication verification, and maintenance. In most situations, interfaces that combine interactive verification, actionable uncertainty communication, selective transparency, and support for intermediate reasoning were more effective than static explanation designs; however, deployment remains constrained by methodological heterogeneity, limited real-world integration, small samples, limited real-world integration and inconsistent measures. This review proposes a thematic structure linking deliberative support, oversight-preserving design, and calibrated transparency, and offers a roadmap for embedding trustworthy human–AI interfaces in safety-critical decision support systems.

Keywords: Human–AI interface; high-stakes decision; cognitive workload; trust calibration.

1. Introduction

Artificial intelligence (AI) is transitioning beyond experimental development into decision support in medicine, transport, and other safety-driven areas, where mistakes might adversely affect patients, jeopardize operations, or endanger populations. The use of AI has become a major part of making decision-making better in various fields, which helps with efficiency and effectiveness (Choi et al., 2022; Alhatimi Aleessawi & Djaghroui, 2025). Through the utilization of machine learning, natural language processing, and neural networks, AI systems possess the capability to analyze extensive datasets, detect patterns, and generate insights to support decision-making. The integration of AI in decision-making offers numerous benefits, including enhanced efficiency, improved accuracy, and advanced predictive analytics (Balbaa & Abdurashidova, 2024). It is interesting due to the capability to transform highly complicated, large-volume data and provide fast predictions or guidance that can promote diagnosis accuracy, efficiency, and early detection of risks (Topol, 2019; Perez-Cerrolaza et al., 2024; Khalifa & Albadaawy, 2024). Prediction is not only a characteristic of high-stakes work. It demands professionals to make evidence-based interpretations, calculate the degree of uncertainty, balance evidence against contradicting information, and make actions and choices that are institutionally, ethically, and legally accountable. This is why the question is no longer about the ability of AI to do well in terms of computation, but about whether the system can fit into a decision environment, being safe, intelligible, and operationally useful (Labkoff et al., 2024; van Baalen et al., 2021). The issue is exacerbated by the fact that most of the AI tools are deployed to overloaded systems. In intensive care, as an example, monitoring is intensive, alarm-saturated, and cognitively demanding. Employees will have to control several devices, give preference to conflicting signals, operate under time restrictions, and contend with ease-of-use issues, visualisation constraints, and poor alarm culture (Poncette et al., 2019; Mosch et al., 2024). As a new support tool is brought to such a place, no attention is paid to it, and the extra complication will only serve to harm the quality of decisions, not enhance it. The combination of AI and proven safety engineering processes is an unsolved problem in industrial and transportation contexts, especially in the areas where human control should still coexist with more autonomous operations (Perez-Cerrolaza et al., 2024). The study of agent transparency in safety-critical work demonstrates that the connection between disclosure of information, situation awareness, workload, and operator performance is consequential and contextual, in particular, when agent transparency is not integrated into the main task interface (van de Merwe et al., 2024).

Another contradiction is that technical openness does not necessarily translate into human comprehension or warranted trust. AI explanations are typically based on technical intuition as opposed to how individuals really create, assess, and utilise explanations in social and cognitive contexts (Miller, 2019). Explainability has thus far been suggested as a pathway to credible AI in the field of healthcare, though its advantages are conditional and partial. Markus et al. (2021) describe that explainability should correspond to the purpose of its requirement, and Ghassemi et al. (2021) warn that the existing explainability techniques cannot be seen as adequate measures that protect decision-support at the patient level. This puts questions of implementation at the centre stage. Safe adoption will rely not just on the claims of model performance but also on selection, acceptance testing, commissioning, training, and quality assurance (Mahadevaiah et al., 2020). It also relies on minimising unnecessary loads as well as enhancing record-keeping, validation, surveillance, and governance throughout the life cycle of AI-powered clinical decision support (Graafsma et al., 2024; Labkoff et al., 2024). Collectively, the literature presents AI-based high-stakes decision support as a sociotechnical issue of interface, workflow, responsibility, and oversight as opposed to a solitary algorithmic issue, hence warranting a detailed overview of the human-AI interface design in the context of consequential decision making and uncertainty.

1.1 Aim and Research Questions

The aim is to map how human–AI interface design influences cognitive workload, trust calibration, and safe use of AI-supported decision systems in high-stakes settings, and to identify the interface features that appear most helpful for effective decision quality and human oversight.

Research questions:

- i. What interface features are being used in high-stakes human–AI systems?
- ii. How do these interface designs affect trust, reliance, workload, and performance?
- iii. What design patterns help prevent over-reliance and support safer human oversight?

2. Methodology

2.1 Design and Reporting Standards

This scoping review employed the Population–Concept–Context (PCC) framework to map the evaluative patterns, characteristics, and breadth of human–AI interface design for trust calibration and cognitive workload management in high-stakes decision-making settings. This study was reported in accordance with the PRISMA extension for scoping reviews (PRISMA-ScR), which included the transparent documentation of study selection decisions and use of a flow diagram. We did not undertake any protocol registration. In line with the scoping review's aim to synthesise, classify, and describe evidence rather than estimate pooled intervention effects, no formal risk of bias or critical appraisal was performed. Methodological safeguards included a multi-database search strategy, staged screening, explicit eligibility criteria, and standardised data charting.

2.2 Eligibility Criteria

The inclusive PCC criteria were applied to identify empirical studies in which human decision-makers interacted with AI-supported interfaces in safety-critical environments. Non-empirical, purely technical reports and low-stakes studies were excluded. A comprehensive literature review was conducted with rigorous screening to ensure that the included studies strictly adhered to the study design criteria and predetermined PCC. This review included studies that had been checked by other scientists and that looked at how people and AI can work together in important areas like healthcare, critical infrastructure, aviation, security operations, and emergency management. Measurable outcomes related to cognitive workload, trust, task performance, reliance, usability, or comparable human factors variables had to be reported in the studies. Table 1 details the inclusion and exclusion criteria applied in this study.

Table 1. Eligibility criteria according to the PCC framework

Item	Inclusion criteria	Exclusion criteria
Population	Human decision-makers in high-stakes settings, including clinicians, pharmacists, therapists, air traffic controllers, power-grid operators, emergency personnel, security analysts, and similar professional users	Non-human studies; animal or simulation-only technical studies without human participants
Concept	Human–AI interface design features such as explanations, confidence displays, uncertainty visualizations, rationale displays, counterfactuals, what-if tools, alerting structures, and oversight controls	AI model development studies with no interface or interaction component; general trust papers with no measurable interface evaluation
Context	High-stakes decision-support environments where errors may produce serious clinical, operational, safety, or security consequences	Low-stakes consumer, educational, entertainment, or general productivity settings
Study Designs	Experimental studies, quasi-experimental studies, randomized studies, observational evaluations, mixed-methods studies, design evaluations, and implementation studies	Narrative reviews, systematic reviews, scoping reviews, editorials, letters, commentaries, protocols without results, and conference abstracts without full text

Item	Inclusion criteria	Exclusion criteria
Outcomes	Any of: trust, reliance, trust calibration, cognitive workload, NASA-TLX, usability, task performance, accuracy, response time, error detection, alert handling, or oversight behavior	Studies not reporting user-centered, behavioral, or performance-related outcomes
Publication Type	Peer-reviewed journal articles and full peer-reviewed conference papers	Preprints, theses, reports, guidelines, books/chapters, and other grey literature
Language	English	Non-English
Timeframe	2010-2025	Studies published outside this timeframe

2.3 Information Sources

The retrieval process was restricted to peer-reviewed literature indexed in selected databases that cover human–computer interaction, human factors, decision support, artificial intelligence, and safety-critical systems particularly well. The following bibliographic databases were consulted: Web of Science, IEEE Xplore, Scopus, PubMed/MEDLINE, the ACM Digital Library, and ScienceDirect. Consultation of grey literature sources was not undertaken. All retrieved records were exported to reference management software for de-duplication. This was followed by staged screening at the title/abstract and full-text levels, applying the predefined eligibility criteria.

2.4 Search Strategy

A series of database-specific search strategies were developed by combining controlled vocabulary terms with free-text keywords covering the PCC elements of the review, where available. The search strategy prioritised sensitivity over specificity to capture the broad range of relevant interface studies across domains. The search terms were centred on four key concept categories: interface or explanation design, trust or workload outcomes, human–AI interaction, and high-stakes contexts. Terms included combinations of: 'AI-assisted decision making', 'human–AI interface', 'automation bias', 'human–AI interaction', 'decision support', 'over-reliance', 'explainable AI', 'NASA-TLX', 'confidence display', 'uncertainty visualisation', 'trust calibration', 'cognitive workload', 'mental workload', 'healthcare', 'aviation', 'emergency management', 'critical infrastructure', and 'security'. Filters were applied where appropriate to restrict the results to peer-reviewed, English-language studies focusing on humans published between 2015 and 2025. Table 2 illustrates the search logic applied across databases.

Table 2. Search strategy structure

Database	Search string
Scopus	(“human-AI interaction” OR “human-AI interface” OR “AI-assisted decision making” OR “decision support”) AND (explain* OR “confidence display” OR uncertainty OR “counterfactual explanation” OR transparency OR oversight) AND (trust OR reliance OR “trust calibration” OR workload OR “mental workload” OR “NASA-TLX”) AND (healthcare OR aviation OR “critical infrastructure” OR security OR “emergency management”)
Web of Science	TS=((“human-AI interaction” OR “decision support”) AND (explain* OR uncertainty OR transparency OR “confidence display”) AND (trust OR reliance OR workload OR “mental workload”) AND (healthcare OR aviation OR security OR “critical infrastructure”))
IEEE Xplore	(“All Metadata”: “human-AI interaction” OR “AI-assisted decision making”) AND (“trust calibration” OR workload OR “automation bias” OR transparency) AND (“decision support” OR healthcare OR aviation OR infrastructure)
ACM Digital Library	(“human-AI interface” OR “AI-assisted decision-making”) AND (explanation OR confidence OR uncertainty OR transparency) AND (trust OR reliance OR workload)
PubMed/MEDLINE	(“Decision Support Systems, Clinical” OR “Artificial Intelligence” OR “Human-Computer Interaction”) AND (trust OR reliance OR workload OR usability) AND (explanation OR uncertainty OR transparency) AND (high-stakes OR healthcare OR patient safety)
ScienceDirect	(“human-AI interaction” OR “AI decision support”) AND (trust OR workload OR reliance) AND (explanation OR transparency OR uncertainty) AND (healthcare OR aviation OR infrastructure OR security)

2.5 Study Selection

All the records that were retrieved were screened in two stages. The records were first filtered in accordance with the predetermined eligibility criteria, which excluded obviously irrelevant ones. Secondly, a full-text evaluation was done on studies that had been selected as potentially relevant to qualify them as final subjects to be included. Only studies that examine HAI design in high-stakes contexts, and provide empirical results concerning cognitive workload, trust, task or reliance performance were retained. The PRISMA-style flow diagram described the study selection process and provided an overview of the number of records identified, screened and excluded, including the final version in the evidence base.

2.6 Data Extraction and Synthesis

The data were charted on a standardised extraction template, which was then refined so as to achieve consistency across the studies that were included. Variables were then identified, including bibliographic, sphere, and decision context, attributes of participants, AI assignment, interface or clarification approach, study blueprint, comparator state, confidence, and workload, performance outcomes, and key discoveries. Additional information about failure modes, including automation over- trust, bias, under-trust, oversight dilemma, alert fatigue, etc. was collected as well where relevant.

The synthesis was mapping and descriptive, as opposed to effect-estimating. The included studies resulted in a numerical summary that included domains of application, years of publication, types of interface strategies, study designs and groups of participants. Further, narrative thematic synthesis was used to classify the results into broad themes that matched with the review questions. These themes discussed the recurrent failure modes coupled with related interface mitigation, workload and performance support, transparency and trust calibration.

3. Results

3.1 Screening and Selection

The database search yielded 436 records. After removing 228 duplicate records, the remaining 208 unique records were screened against the above criteria using the title and abstract, and 85 papers were retained for full-text assessment. One hundred and twenty-three records were excluded at the title–abstract stage because they did not match the scope of the review. A further 68 papers were excluded during the full-text review stage: 28 due to being editorials, conference abstracts without full text, reviews, preprints, protocols, theses, reports, guidelines, books/chapters and other grey literature; 16 due to being AI model development studies with no interface or interaction component, general trust papers with no measurable interface evaluation and low-stakes settings; and 14 due to being non-human or simulation-only technical studies without human participants. Finally, 17 studies that met all the criteria were synthesised (see Table 3 and Fig. 1).

3.2 Summary of Study Characteristics

The evidence base included 17 empirical studies extracted from the uploaded study pack, and this was published safe reliance, within the clinical decision support, medical imaging, sepsis management, telehealth diagnosis, medication check, rehabilitation evaluation, and tumour interpretation. Fewer were related to air traffic control, Power-grid congestion management, and condition-based maintenance, which made the review have a health-centric, albeit cross-domain profile. The studies were heterogeneous methodologically and encompassed randomised experiments, controlled laboratory studies, mixed-methods evaluations, factorial experiments, scenario-based factorial experiments, heuristic evaluations, user-centred design studies, and case studies. Sample sizes were varied, with 6 clinical experts to 292 participants; participants were physicians, pharmacists, pathologists, therapists, air traffic controllers, power-grid operators, and novices. The interface strategies that were tested were explanations, confidence or uncertainty displays, counterfactuals, visual analytics, interpretable recommendations, refinement tools, and workflow integrations. Some of the common outcomes include trust, reliance, workload, usability, reaction time, agreement, and task performance. The conceptual pathway through which core human–AI interface features influence contestable oversight, deliberative support, calibrated transparency, and downstream decision outcomes in high-stakes settings is summarised by Fig. 2.

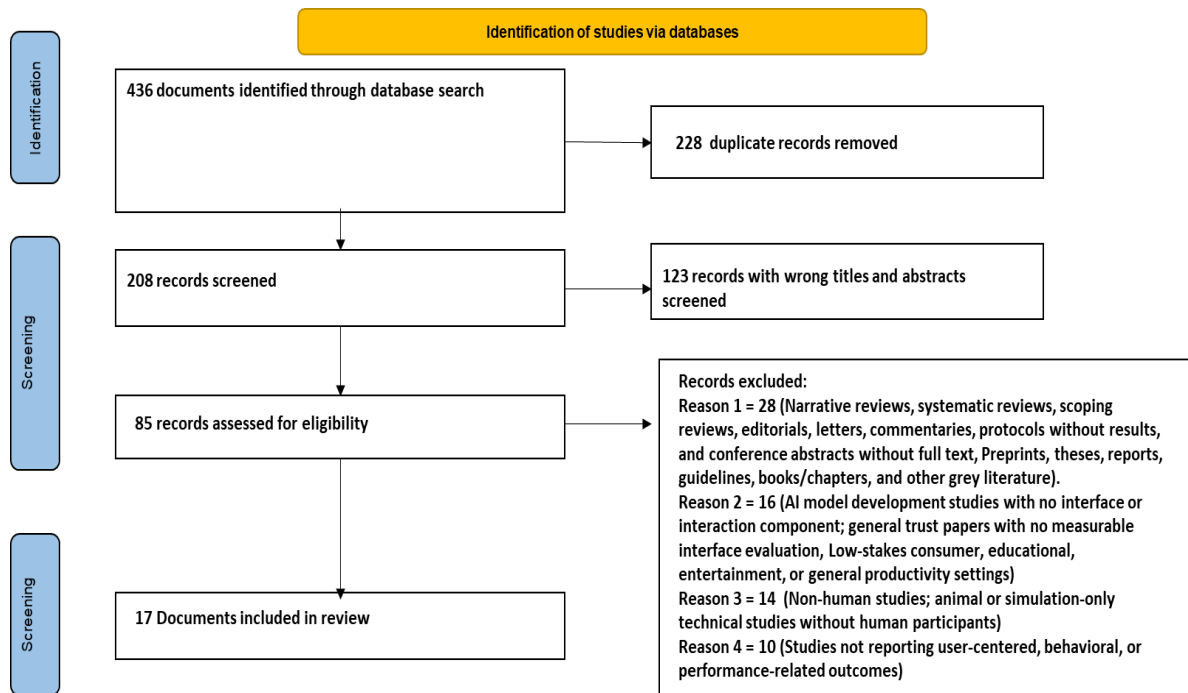


Fig. 1. PRISMA flow diagram

A synthesis of the manuscript's three results themes

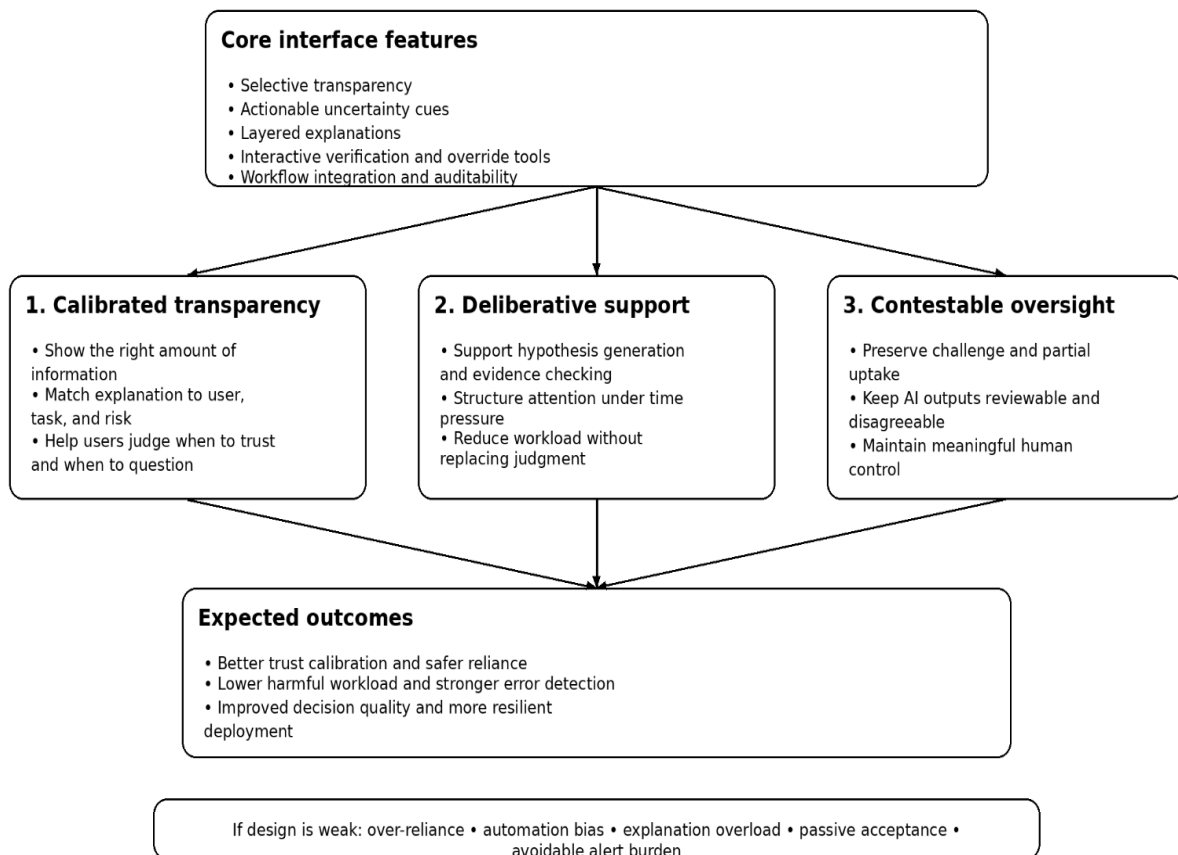


Fig. 2. Conceptual framework for high-stakes human-AI interface design

Table 3. Data extraction table for the included studies

Study	High-stakes context/ participants	AI task/ decision setting	HAI interface or interaction strategy	Study design/ comparator	Trust, workload, and performance measures	Key findings	Failure mode/design implication
Ackerhans et al. (2025) – healthcare/ sepsis CDSS	Hospital sepsis prevention scenario; 292 medical students and physicians across specialties.	AI-based CDSS for sepsis prediction/prevention.	Manipulated explainability of decision outcome, depth of workflow integration, and accountability/signature requirement.	Web-based 2×2 factorial between-subject scenario experiment.	Trust in AI; perceived professional identity threat; process-design effects (no workload metric reported in uploaded extract).	Explainability increased trust ($\beta=.508$, $P<.001$) but also increased identity threat ($\beta=.351$, $P=.02$). Deeper workflow integration increased trust ($\beta=.262$, $P=.009$). Trust partially reduced identity threat ($\beta=-.138$, $P=.047$). Accountability/signature increased identity threat ($\beta=.339$, $P=.004$).	Transparency and integration can strengthen trust, but accountability and visible AI intrusion into professional roles can trigger resistance; trust calibration must be balanced with role protection.
Bussone et al. (2015) – clinical diagnosis CDSS	Healthcare professionals diagnosing hypothetical balance-disorder cases using fictional patients.	Prototype CDSS for diagnostic support.	Compared fuller versus less-detailed explanations; also varied displayed system confidence.	Exploratory user study with observation of trust and reliance behavior.	Trust, reliance/over-reliance, self-reliance, perceived usefulness of confidence/explanations.	Displayed confidence had only slight effects. Fuller explanations increased trust but also promoted over-reliance. Less-detailed explanations reduced trust and led to self-reliance problems.	More explanation is not automatically safer; explanation depth must help clinicians verify reasoning, interpret confidence, and make differential diagnoses rather than merely persuade.
Cai et al. (2019) – pathology image retrieval	Pathologists using ML-based image retrieval during medical decision-making; two evaluations (sample size not stated in uploaded extract).	Deep-learning retrieval of visually similar pathology images from prior cases.	Human-centered refinement tools enabling on-the-fly control of similarity criteria and exploration of retrieved cases.	Needs assessment plus two interface evaluations against a more traditional interface.	Diagnostic utility of retrieved images, user trust, interface preference, diagnostic accuracy.	Refinement tools increased diagnostic utility and user trust, were preferred over a traditional interface, and did not reduce diagnostic accuracy. Users also repurposed the tools to interrogate the algorithm and distinguish model error from human error.	Interactive controls can support calibrated trust by making algorithmic similarity contestable and tunable, reducing passive acceptance of retrieved outputs.

Study	High-stakes context/ participants	AI task/ decision setting	HAI interface or interaction strategy	Study design/ comparator	Trust, workload, and performance measures	Key findings	Failure mode/design implication
Gentile et al. (2025) – condition-based maintenance	Hydraulic-system condition diagnosis; 24 participants in a digital work environment.	Supervised ML decision aid for maintenance diagnosis.	Compared baseline/no explanation, normative+contrastive explanations, normative+counterfactual explanations, and all three combined.	Controlled experiment with four explanation conditions.	Decision performance, false alarm rate, decision time, workload.	No significant overall performance differences across explanation conditions. The combined normative+contrastive+counterfactual condition reduced false alarms, while counterfactual-containing conditions showed potential reductions in decision time and workload.	Counterfactuals may ease cognitive burden, but their value depends on how they are bundled with other explanation forms; explanation complexity should not be assumed to improve accuracy.
Gomez et al. (2024) – telehealth infectious-disease screening	Telehealth strep throat screening; 121 telehealth providers.	AI CDSS predicting streptococcal pharyngitis from smartphone throat images.	Three explainable-AI CDSS prototypes introduced after a validated baseline aid (Modified Centor Score).	Randomized online experiment using clinical vignettes and linear-model analysis.	Prediction accuracy, confirmatory testing rate, perceived trust, perceived understanding, agreement with AI.	AI-based CDSS improved clinicians' predictions relative to the Centor Score. Yet providers reported lower trust in AI advice than in the Centor baseline and requested more in-person confirmatory testing despite higher agreement with AI outputs.	Accuracy gains do not guarantee calibrated acceptance; unfamiliar AI representations may provoke defensive checking, especially where overtreatment and missed diagnosis carry consequences.
Hurter et al. (2024) – air traffic control	21 air traffic controllers evaluating a machine-learning DSS in air traffic management.	ML-supported conflict management / decision support in ATM.	Compared three levels of explainability and visualization transparency in the DSS.	Controlled validation study/case study with end-user air traffic controllers.	User ratings of transparency, usefulness, acceptance, training value, and trust-related perceptions.	Explainable and transparent AI was seen as promising for trust-building, training, and operational acceptance, with benefits strongest when information improved controller understanding rather than merely adding detail.	In safety-critical control rooms, transparency must support controller reasoning and training, not overload already complex displays.

Study	High-stakes context/ participants	AI task/ decision setting	HAI interface or interaction strategy	Study design/ comparator	Trust, workload, and performance measures	Key findings	Failure mode/design implication
Lee & Chew (2023) – post-stroke motion assessment	7 physical therapists and 10 non-experts evaluating AI-based motion assessment.	Clinical assessment support for post-stroke movement quality.	Compared counterfactual explanations with saliency-map explanations.	Experimental comparison of explanation formats under correct and incorrect AI predictions.	Reliance/over-reliance, correctness of final decisions, explanation preference.	Even when aware of AI fallibility, users over-relied on incorrect AI outputs. Counterfactual explanations reduced over-reliance by 21% compared with saliency explanations.	Counterfactuals appear more effective than feature-highlighting alone for helping users challenge wrong AI decisions.
Lester et al. (2025) – medication verification	30 pharmacists in a web-based mock medication verification task.	AI support for medication verification and approval decisions.	Compared black-box AI versus uncertainty-aware AI.	Randomized controlled experiment.	Reaction time, rate of harmful approval decisions, trust-related safe reliance behavior.	Uncertainty-aware AI reduced reaction times in most cases and helped pharmacists avoid harmful incorrect approvals compared with black-box AI.	Communicated uncertainty can improve both speed and safety when the task is verification-focused and time sensitive.
Naiseh et al. (2023) – clinical XAI interfaces	Clinical decision-support context; participant details not fully specified in uploaded extract.	Clinical AI decision support.	Compared four explanation classes for XAI interfaces.	Empirical comparison of explanation classes.	Trust calibration, understandability, explanation usefulness.	Two explanation classes were more understandable than the others, but explanations could still trigger over-reliance depending on presentation and context.	Explanation type matters; interpretability must be aligned with the user's task and not treated as inherently trust-calibrating.
Prince et al. (2025) – pediatric neuro-oncology imaging	6 expert clinicians interpreting radiographic images of adamantinomatous craniopharyngioma.	AI-assisted radiographic interpretation.	AI decision-support overlays/recommendations integrated into expert image interpretation workflow.	Applied expert evaluation/case study.	Perceived difficulty, efficiency, confidence, expert agreement/disagreement.	AI support reduced perceived difficulty in most cases and was viewed as useful for efficiency, but it also surfaced disagreement and interpretive tension among experts.	Even high-expertise users may not converge around AI outputs; interfaces should preserve room for expert deliberation rather than imply singular correctness.
Salimzadeh et al. (2024) – uncertainty/reliance study	Experimental assisted decision-making task; participant details	Prognostic versus diagnostic AI-assisted decision-making under	Manipulated task framing and uncertainty conditions.	Comparative experimental study.	Trust, reliance, task difficulty effects.	Increasing task complexity and uncertainty increased reliance on AI without	Reliance and trust are not interchangeable; interface evaluation should capture both

Study	High-stakes context/ participants	AI task/ decision setting	HAI interface or interaction strategy	Study design/ comparator	Trust, workload, and performance measures	Key findings	Failure mode/design implication
(scope-borderline)	not fully specified in uploaded extract.	uncertainty.				necessarily increasing self-reported trust.	behavioral uptake and subjective trust.
Sivaraman et al. (2023) – sepsis treatment recommendations	24 intensive care clinicians using AI-based sepsis treatment recommendations.	AI recommendations for sepsis management.	Interpretable recommendation interface showing treatment advice and explanation elements.	Qualitative/experimental evaluation of clinician interaction with AI recommendations.	Acceptance, negotiation behavior, confidence in recommendations, interpretability.	Clinicians did not simply accept or reject recommendations; they often negotiated with them by modifying, delaying, or partially using the advice. Interpretability increased confidence but did not eliminate selective uptake.	Effective HAI may need to support negotiation and partial adoption, not just binary acceptance/rejection.
Vicente & Matute (2023) – AI bias inheritance	Simulated medical diagnosis task with human participants.	AI-assisted diagnostic classification.	Participants exposed to biased AI recommendations during diagnostic decisions.	Experimental study of bias transfer and persistence after AI removal.	Decision bias, persistence of biased judgments, trust-related behavioral carryover.	Participants inherited the AI's diagnostic bias and continued reproducing it even after the AI support was removed.	Automation effects can persist beyond immediate interaction; interfaces should actively help users detect and correct biased guidance.
Viebahn et al. (2025) – power grid congestion management	Full end-user population of a transmission-system-operator DSS in a real power-grid management context.	AI-based decision support for congestion management.	Operational DSS evaluated for UX, trust, and workload in grid-control setting.	Applied user-experience evaluation with end users.	Trust/acceptance, task success, NASA-TLX workload, System Usability Scale (SUS).	The DSS received high overall UX scores with acceptable workload and good usability, demonstrating the value of formal UX evaluation in critical infrastructure deployment.	Real-world deployment should assess workload and usability alongside technical performance to avoid hidden operational burdens.
Xie et al. (2020) – CheXplain chest X-ray support	Physicians interacting with AI-supported chest X-ray analysis; survey, co-design, and evaluation phases.	AI-enabled medical imaging analysis for chest radiographs.	Co-designed explanation interface enabling exploration and understanding of model outputs.	User-centered design study with clinician involvement across design and evaluation.	Perceived usefulness, interpretability, support for reasoning and exploration.	Physicians valued interfaces that allowed exploration, justification, and understanding of outputs rather than static one-shot explanations.	Explanation interfaces should support sense-making and inquiry, especially in image-based diagnosis where users need to inspect evidence paths.

Study	High-stakes context/ participants	AI task/ decision setting	HAI interface or interaction strategy	Study design/ comparator	Trust, workload, and performance measures	Key findings	Failure mode/design implication
Zhang et al. (2024) – SepsisLab	6 clinicians evaluating an uncertainty-aware sepsis-diagnosis support prototype.	Sepsis diagnosis support with future-risk estimation.	Prototype visualized future sepsis risk, uncertainty, and recommended additional laboratory tests.	Heuristic evaluation / prototype assessment with clinicians.	Perceived utility, support for uncertainty handling, diagnostic reasoning support.	Clinicians viewed the system as promising because it helped them reason about uncertainty and identify next steps, rather than simply presenting a final risk score.	Uncertainty is more actionable when paired with guidance on what information could reduce it.
Zohrevandi et al. (2022) – ecological visual analytics in ATC	45 novices plus expert comparison in air traffic control conflict-resolution tasks.	Visual analytics support for ATC conflict understanding and resolution.	Ecological interface/visual analytics emphasizing system relationships and constraints.	Comparative human-in-the-loop study.	Problem-solving performance, understanding of airspace conflicts, expert–novice comparison.	Ecological visual analytics improved understanding of conflicts and problem-solving performance, with gains driven more by representational support than by prior expertise alone.	Well-designed visual representations can reduce cognitive burden and support reasoning in complex, time-critical control environments.

3.3 Thematic Analysis

Theme 1: Precision transparency: Why more explanation is not always better

In the studies that were included, transparency was most effective when customised to the role of the user, to the task, and decision pressure, as opposed to being presented maximally or consistently. Bussone et al. (2015) discovered that fuller explanations enhanced user trust, and they also motivated over-reliance, but thinner explanations prompted the user to engage in self-reliance in clinical decision support. This trend implies that the degree of explanation is not an adequate guarantee of proper trust. An identical tension was reported by Ackerhans et al. (2025), who found that explainability, together with greater workflow integration, enhanced trust in AI-based sepsis support, but explainability also served as a cause of perceived professional identity threat, and explicit mention of accountability escalated that conflict. Naiseh et al. (2023) also demonstrated that the effect of explanation varied across explanation classes, which shows that the usefulness of transparency might be reliant on how appropriately this practice fits clinical requirements.

Xie et al. (2020) demonstrated that physicians favoured explanatory interfaces that belonged to exploration, justification, and role-sensitive types of reasoning over a one-size-fits-all presentation. Cai et al. (2019) also observed that refinement tools enhanced trust because pathologists were able to reshape algorithmic similarity in context rather than passively receive retrieved cases. The same was the case with uncertainty displays. Lester et al. (2025) demonstrated that uncertainty-conscious AI had led to a safer and faster process of verification of medications, whereas Zhang et al. (2024) established that uncertainty became utilizable when accompanied by recommending next actions, e.g., recommended laboratory testing methods. Conversely, Hurter et al. (2024) have shown that in air traffic control, the mere use of additional explanation was not necessarily helpful, unless it had a direct contribution to operational understanding. Collectively, these studies indicate that the key to effective transparency does not lie in displaying additional information, but in displaying the appropriate information in the right form that assists users to trust, to doubt, and to take action.

Theme 2: Supporting deliberation: Interfaces that improve workload and performance

Human-AI interfaces enhanced performance and assisted in workload management when they facilitated the intermediate reasoning of the users as opposed to merely giving the final recommendation. This was especially clear in the diagnosis of sepsis: Zhang et al. (2024) claimed that a previous tool had been abandoned because it intervened at the very end of the decision-making process; however, SepsisLab assisted clinicians in making earlier assumptions, examining ambiguity, and finding further tests that may decrease it. Cai et al. (2019) demonstrated a similar pathological process. Their refinement tools not only described the retrieval model, but also enabled pathologists to navigate the search mechanism, test hypotheses, and provide a better opportunity to control the model error against their inherent uncertainty, thus enhancing diagnostic utility without compromising accuracy.

Regarding air traffic control, Zohrevandi et al. (2022) found that the ecological visual analytics enhanced the user comprehensiveness of airspace conflicts and problem-solving performance, and the benefits were associated with the representational support rather than with previous expertise. Hurter et al. (2024) also demonstrated that explainable support in air traffic management was appreciated when the support enhanced comprehension and the utility of training. According to Prince et al. (2025), the use of AI aid made tasks less difficult and more efficient in specialist radiography interpretation, yet it also created a conflict within the circles of specialists. Lester et al. (2025) offered one of the most transparent workload results, demonstrating that the use of uncertainty-aware AI reduced the reaction time and defended pharmacists against damaging false approvals compared to the black-box AI. Gentile et al. (2025) applied this outside of the healthcare domain, demonstrating that groups of explanations including counterfactuals can lower the number of false alarms, workload, and the time needed to make decisions in condition-based maintenance. In a power-grid setting, Viebahn et al. (2025) then presented the idea that trust, usability, workload, and task performance could be collectively evaluated to provide a workable measure of operational deployment. In general, the finding suggests the effective interface can enhance results by providing structure to attention, verification, and reasoning instead of merely increasing the information displayed on the screen.

Theme 3: Guardrails against failure: Why safe use depends on contestability and oversight

The third theme highlighted that even in systems that seemed to be a success, failure modes still existed, suggesting the need to actively build in the role of humans' oversight instead of assuming it existed. The most

prominent risk was identified as over-reliance. As shown by Bussone et al. (2015), more elaborate explanations might unwillingly lead clinicians to follow flawed recommendations. On the contrary, Lee and Chew (2023) also found out that the users still relied heavily on erroneous AI outputs even though they were aware that the system was fallible. Counterfactual explanations did not eliminate the over-reliance, but reduced it by 21%. A more profound issue was discovered by Vicente & Matute (2023): the users were influenced by biased AI suggestions greatly, which caused them to repeat mistakes made whenever the AI was used, even after its withdrawal. This means that problematic dependency may persist even after the immediate engagement.

Other researchers have demonstrated that failure with regard to trust is not simply about blind acceptance. Salimzadeh et al. (2024) provided an important behavioural lesson by demonstrating that an increase in uncertainty and complexity led to an increase in reliance with no corresponding rise in reported trust, but this was more of a borderline case when compared to the scope of the review. This shows that self-reported trust is an incompetent measure of safe use. Sivaraman et al. (2023) also found that ICU clinicians did not immediately endorse sepsis recommendations; rather, they negotiated with them by modifying, postponing or partially adopting the advice. Another kind of guarded supervision was recorded by Gomez et al. (2024): healthcare providers were more willing to agree with the AI-generated results, but they still demanded further confirmation testing due to a lack of confidence relative to effectiveness. The research by Ackerhans et al. (2025) and Prince et al. (2025) also demonstrated that AI is capable of causing interpretive tension and disagreement, as well as identity threat, even in cases of efficiency improvement. Taken together, these studies imply that a more secure human-AI system is one that does not eradicate contestability, allows partial adoption, and brings out the existence of standardised biases.

4. Summary of Key Findings

In the studies covered by the study, the key finding was that the meaningful impact of the interface design on the effect of AI on improving judgment or promoting harmful dependency was large in all reviewed articles. Explanations also enhanced trust, but comprehensive or inappropriately matched explanations might also enhance over-reliance, whereas shallow explanations may result in distrust or self-reliance. The most promising strategies were counterfactual and interactive forms of explanations, in particular, to assist users in questioning AI-based recommendations and diminishing the acceptance of erroneous outputs. Uncertainty-aware interfaces tended to perform better than black-box advice, such as by reducing reaction time, protecting against erroneous advice, or encouraging the user to be more cautious, but tended to prompt additional checking. Ecological visual analytics and interpretable decision-support displays showed more understanding or problem solving or workflow fit in operational situations. Nevertheless, the facts have revealed some persistent failure modes, including: automation bias, inherited AI bias, threats to professional identity, disagreement among experts, and biased but not exhaustive acceptance of recommendations. Altogether, the best interfaces were those that enabled verification, contestability, and intermediate reasoning, and not active acceptance of AI results.

5. Comparison with Global Literature

The present review showed that human-AI collaboration can be improved by transparency, but only when it is actionable, role-sensitive, and selective, and not maximal. According to Hoff and Bashir (2015), trust is stratified and is influenced by learned, situational, and dispositional factors. This provides an explanation as to why the same explanation can calibrate trust in one situation and distort it in another. Similarly to Mehrotra et al. (2024), Wischniewski et al. (2023) also indicate that the concept of trust calibration is multidimensional and that improper measurement often misleads about whether transparency improves proper reliance or just increases subjective trust. The findings also correspond to those of Zhang et al. (2020), who have discovered that confidence cues can enhance trust calibration even without a clear enhancement of joint accuracy. Equally, Yu and Li (2022) demonstrated that transparency may boost the perceived discomfort and effectiveness, which supports the notion that useful transparency should be context-related.

Interfaces supporting intermediate reasoning as opposed to simply offering final recommendations have been observed to facilitate performance and assist in managing workload. This conclusion follows a critique within the field of healthcare human factors, where AI studies tend to overlook workflow, cognition, and ecological validity (Asan & Choudhury, 2021). It is also in accordance with the Measurement–Algorithm–Presentation model suggested by van Berkel et al. (2023), which argues that safety in decision support is rooted in how output is given to be interpreted. According to Bansal et al. (2021), explanations led to more accepting AI

advice, yet no improvement in complementary team performance. Meanwhile, Bućinca et al. (2021) have shown that such reduction of overreliance often requires cognitive forcing and not explanation alone, but usability problems can also accompany this method. Moreover, Cao and Huang (2022) also define the difference between trust and reliance, implying that behavioural measures like gaze can be more effective at workload measurement than self-reports.

The third theme was persistence of failure modes, especially over-reliance on automation, bias, and contestability required. Dratsch et al. (2023) have shown that radiologists could be prone to inaccurate AI suggestions, and Foroughi et al. (2023) showed that highly automated systems changed attention and decreased discovery failures, which increases operational risk. A further demonstration of the harm caused by automation failures and of the partial and behaviourally imperative character of trust repair is given by Mishler & Chen (2023) and Centeio Jorge et al. (2023). Carter et al. (2024) support this claim in this review that trust calibration is not as dependent on the superficial, human-like evidence but, rather, on meaningful discourse related to uncertainty. Overall, researchers already attest to the fact that instead of passive compliance, complementary performance depends on the decisive effect of manageable information load, interaction designs, and correct mental models that provide incentives to selective reliance, verification, and challenge (Steyvers & Kumar, 2024; Ling et al., 2024).

6. Implications for Policy and Practice

Policy: It is suggested that requirements in the human–AI interface must be implemented as a formal safety and quality standard in high-stakes decision-support environments. This would entail obligatory human-centred review prior to deployment, regular evaluation of trust calibration and workload impacts, and minimum interface necessities toward uncertainty communication, explanation traceability, as well as override choices and accountability recording. Human supervision also needs to be an operational need and not a symbolic precaution, and the policies must be established such that AI suggestions may be argued with and verified by auditing them at the point of usage. Besides this, the procedures of procurement and regulatory approval must demand evidence of not just model accuracy, but must also necessitate a demonstration of model workflow fit, usability, alert load, and impact on the quality of decision in realistic operating conditions. Audits and post-deployment monitoring should be specified with core evaluation indicators of trust calibration, NASA-TLX, or equivalent measures of workload, accuracy of the tasks performed, ability to detect errors, capacity to handle false alarms, and override behaviour. In resource-constrained environments, simple interface features indicating high value, such as clear confidence or uncertainty indicators, a concise estimation of appropriately specific roles and purposes, and a minimal burden escalation procedure, should be prioritized and facilitated by gradual implementation and specific staff education.

Practice: It is necessary to practice the adoption of a review, verify, decision-monitor model and interface design, facilitating users to verify AI outputs, quantify uncertainty, demand a comparison with evidence, case-specific, and record final decisions. Implementation of stratified explanations is also essential, and brief actionable summaries are displayed as default, and a deeper resort to rationale, counterfactuals, or supporting evidence on request. To promote safe uptake, interfaces are suggested to have verification supports like secondary checks, prioritisation of alerts, and easy-access overrides, and, in the absence of overload in interface visual presentation, generic explanation overload, or instant persuasion indications. On the other hand, embedding AI outputs into workflows should be avoided in such a manner that can be seen as creating passive acceptance or making disagreement challenging. Enhancement of interface behaviour with respect to complex tasks, expertise of the user, time sensitivity, and risk of the domain are requirements that must be customised. Trust, reliance, workload, response time, and error-detection metrics must be used to monitor the results of use, and the resulting information must be used in cyclic redesign processes. It is advised that the training of the teams on the interpretation of uncertainty, proper reliance, and the practice of escalation is strengthened and practiced to enhance the safety, adoption, and equity in reality settings.

7. Conclusion

This scoping review synthesised 17 empirical studies on human–AI interface design for cognitive workload and trust calibration management in high-stakes decision-making contexts. Across diverse domains such as healthcare, maintenance settings, power grid operations, and air traffic control, interfaces combining selective transparency, interactive verification, actionable uncertainty communication, and user-centred reasoning

generally showed more promise than static, one-size-fits-all explanation designs. However, persistent failure modes such as guarded uptake, identity threat, bias inheritance, and over-reliance remained evident. However, most studies were methodologically heterogeneous and small, with limited standardisation in trust and workload measurement, and only loosely connected to sustained real-world deployment. Overall, the evidence suggests that human–AI systems can manage cognitive demand and improve decision quality when interface features support context-sensitive human oversight, verification, and contestability, rather than passive acceptance of model outputs. Future work should prioritise stronger cross-domain benchmarking, external validation of interface effects, ecologically valid field evaluation, and more consistent use of behavioural and workload measures alongside subjective trust ratings. There should also be a focus on developing explainable, integrated, uncertainty-aware tools linked to clear performance indicators such as override behaviour, response time, error detection, and safe reliance. These gaps must be addressed if high-stakes human–AI decision support is to evolve from promising prototypes into a trustworthy, mature, and operationally resilient capability.

8. Limitations of Review

Despite the provision of a structured cross-domain map, this review has limitations. The evidence base was small (17 studies), often based on simulations, prototypes, methodologically heterogeneous, healthcare-dominant, and or scenario experiments rather than long-term real-world deployment. Furthermore, trust and workload were measured inconsistently, which limited direct comparison across studies. No formal risk-of-bias appraisal was undertaken. Only peer-reviewed studies written in English were included. Grey literature was also excluded, potentially omitting implementation insights from operational settings and reducing geographic and sectoral representativeness.

Disclaimer (Artificial Intelligence)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

Competing Interests

Authors have declared that no competing interests exist.

References

- Ackerhans, S., Wehkamp, K., Petzina, R., Dumitrescu, D., & Schultz, C. (2025). Perceived trust and professional identity threat in AI-based clinical decision support systems: Scenario-based experimental study on AI process design features. *JMIR Formative Research*, 9, e64266. <https://doi.org/10.2196/64266>
- Alhatimi Aleessawi, N. A. K., & Djaghroui, L. (2025). Artificial intelligence in decision-making: Literature review. *Journal of the Association of Arab Universities for Higher Education Research*, 45(1).
- Asan, O., & Choudhury, A. (2021). Research trends in artificial intelligence applications in human factors health care: Mapping review. *JMIR Human Factors*, 8(2), e28236. <https://doi.org/10.2196/28236>
- Balbua, M. E., & Abdurashidova, M. S. (2024). The impact of artificial intelligence in decision making: A comprehensive review. *EPRA International Journal of Economics, Business and Management Studies*, 11(2), 27–38.
- Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., ... & Weld, D. (2021, May). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). <https://doi.org/10.1145/3411764.3445717>
- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–21. <https://doi.org/10.1145/3449287>
- Bussone, A., Stumpf, S., & O’Sullivan, D. (2015). The role of explanations on trust and reliance in clinical decision support systems. In W.-T. Fu, P. Balakrishnan, S. Harabagiu, F. Wang, & J. Srivatsava (Eds.), *Proceedings – 2015 IEEE International Conference on Healthcare Informatics, ICHI 2015* (pp. 160–169). Article 7349687. Institute of Electrical and Electronics Engineers Inc. <https://doi.org/10.1109/ICHI.2015.26>

- Cai, C. J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., ... & Terry, M. (2019, May). Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). <https://doi.org/10.1145/3290605.3300234>
- Cao, S., & Huang, C. M. (2022). Understanding user reliance on AI in assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), 1–23. <https://doi.org/10.1145/3555572>
- Carter, O. B. J., Loft, S., & Visser, T. A. W. (2024). Meaningful communication but not superficial anthropomorphism facilitates human-automation trust calibration: The human-automation trust expectation model (HATEM). *Human Factors*, 66(11), 2485–2502. <https://doi.org/10.1177/00187208231218156>
- Centeio Jorge, C., Bouman, N. H., Jonker, C. M., & Tielman, M. L. (2023). Exploring the effect of automation failure on the human’s trustworthiness in human-agent teamwork. *Frontiers in Robotics and AI*, 10, 1143723. <https://doi.org/10.3389/frobt.2023.1143723>
- Choi, S., Kim, N., Kim, J., & Kang, H. (2022). How does AI improve human decision-making? Evidence from the AI-powered Go program. *Preprint at SSRN*. <https://doi.org/10.2139/ssrn.3893835>
- Dratsch, T., Chen, X., Rezazade Mehrizi, M., Kloeckner, R., Mähringer-Kunz, A., Püsken, M., Baeßler, B., Sauer, S., Maintz, D., & Pinto Dos Santos, D. (2023). Automation bias in mammography: The impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology*, 307(4), e222176. <https://doi.org/10.1148/radiol.222176>
- Foroughi, C. K., Devlin, S., Pak, R., Brown, N. L., Sibley, C., & Coyne, J. T. (2023). Near-perfect automation: Investigating performance, trust, and visual attention allocation. *Human Factors*, 65(4), 546–561. <https://doi.org/10.1177/00187208211032889>
- Gentile, D., Donmez, B., & Jamieson, G. A. (2025). Human performance effects of combining counterfactual explanations with normative and contrastive explanations in supervised machine learning for automated decision assistance. *International Journal of Human-Computer Studies*, 196, 103434. <https://doi.org/10.1016/j.ijhcs.2024.103434>
- Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11), e745–e750. [https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9)
- Gomez, C., Smith, B. L., Zayas, A., Unberath, M., & Canares, T. (2024). Explainable AI decision support improves accuracy during telehealth strep throat screening. *Communications Medicine*, 4(1), 149. <https://doi.org/10.1038/s43856-024-00568-x>
- Graafsma, J., Murphy, R. M., van de Garde, E. M. W., Karapinar-Çarkit, F., Derijks, H. J., Hoge, R. H. L., Klopotoska, J. E., & van den Bemt, P. M. L. A. (2024). The use of artificial intelligence to optimize medication alerts generated by clinical decision support systems: A scoping review. *Journal of the American Medical Informatics Association: JAMIA*, 31(6), 1411–1422. <https://doi.org/10.1093/jamia/ocae076>
- Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3), 407–434. <https://doi.org/10.1177/0018720814547570>
- Hurter, C., Degas, A., Guibert, A., Poyer, M., Durand, N., Veyrie, A., ... & Aricó, P. (2024). Examining decision-making in air traffic control. Enhancing transparency and decision support through machine learning, explanation, and visualization. A case study. In *International Conference on Agents and Artificial Intelligence* (Vol. 2, pp. 622–634). <https://doi.org/10.5220/0012471900003636>
- Khalifa, M., & Albadawy, M. (2024). AI in diagnostic imaging: Revolutionising accuracy and efficiency. *Computer Methods and Programs in Biomedicine Update*, 5, 100146. <https://doi.org/10.1016/j.cmpbup.2024.100146>
- Labkoff, S., Oladimeji, B., Kannry, J., Solomonides, A., Leftwich, R., Koski, E., Joseph, A. L., Lopez-Gonzalez, M., Fleisher, L. A., Nolen, K., Dutta, S., Levy, D. R., Price, A., Barr, P. J., Hron, J. D., Lin, B., Srivastava, G., Pastor, N., Luque, U. S., Bui, T. T. T., ... Quintana, Y. (2024). Toward a responsible future: Recommendations for AI-enabled clinical decision support. *Journal of the American Medical Informatics Association: JAMIA*, 31(11), 2730–2739. <https://doi.org/10.1093/jamia/ocae209>
- Lee, M. H., & Chew, C. J. (2023). Understanding the effect of counterfactual explanations on trust and reliance on AI for human-AI collaborative clinical decision making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2), 1–22. <https://doi.org/10.1145/3610218>
- Lester, C., Rowell, B., Zheng, Y., Co, Z., Marshall, V., Kim, J. Y., Chen, Q., Kontar, R., & Yang, X. J. (2025). Effect of uncertainty-aware AI models on pharmacists’ reaction time and decision-making in a web-based

- mock medication verification task: Randomized controlled trial. *JMIR Medical Informatics*, 13, e64902. <https://doi.org/10.2196/64902>
- Ling, S., Zhang, Y., & Du, N. (2024). More is not always better: Impacts of AI-generated confidence and explanations in human-automation interaction. *Human Factors*, 66(12), 2606–2620. <https://doi.org/10.1177/00187208241234810>
- Mahadevaiah, G., Rv, P., Bermejo, I., Jaffray, D., Dekker, A., & Wee, L. (2020). Artificial intelligence-based clinical decision support in modern medical physics: Selection, acceptance, commissioning, and quality assurance. *Medical Physics*, 47(5), e228–e235. <https://doi.org/10.1002/mp.13562>
- Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113, 103655. <https://doi.org/10.1016/j.jbi.2020.103655>
- Mehrotra, S., Degachi, C., Vereschak, O., Jonker, C. M., & Tielman, M. L. (2024). A systematic review on fostering appropriate trust in Human-AI interaction: Trends, opportunities and challenges. *ACM Journal on Responsible Computing*, 1(4), 1-45. <https://doi.org/10.1145/3696449>
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- Mishler, S., & Chen, J. (2023). Effect of automation failure type on trust development in driving automation systems. *Applied Ergonomics*, 106, 103913. <https://doi.org/10.1016/j.apergo.2022.103913>
- Mosch, L., Sümer, M., Flint, A. R., Feufel, M., Balzer, F., Mörike, F., & Poncette, A. S. (2024). Alarm management in intensive care: Qualitative triangulation study. *JMIR Human Factors*, 11, e55571. <https://doi.org/10.2196/55571>
- Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2023). How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies*, 169, 102941. <https://doi.org/10.1016/j.ijhcs.2022.102941>
- Perez-Cerrolaza, J., Abella, J., Borg, M., Donzella, C., Cerquides, J., Cazorla, F. J., ... & Flores, J. L. (2024). Artificial intelligence for safety-critical systems in industrial and transportation domains: A survey. *ACM Computing Surveys*, 56(7), 1–40. <https://doi.org/10.1145/3626314>
- Poncette, A. S., Spies, C., Mosch, L., Schieler, M., Weber-Carstens, S., Krampe, H., & Balzer, F. (2019). Clinical requirements of future patient monitoring in the intensive care unit: Qualitative study. *JMIR Medical Informatics*, 7(2), e13064. <https://doi.org/10.2196/13064>
- Prince, E. W., Mirsky, D. M., Hankinson, T. C., & Görg, C. (2025). Impact of AI Decision Support on Clinical Experts' Radiographic Interpretation of Adamantinomatous Craniopharyngioma. *AMIA Annual Symposium proceedings*, 930–939. <https://pubmed.ncbi.nlm.nih.gov/40417524/>
- Salimzadeh, S., He, G., & Gadiraju, U. (2024, May). Dealing with uncertainty: Understanding the impact of prognostic versus diagnostic tasks on trust and reliance in human-AI decision making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1–17). <https://doi.org/10.1145/3613904.3641905>
- Sivaraman, V., Bukowski, L. A., Levin, J., Kahn, J. M., & Perer, A. (2023). Ignore, trust, or negotiate: Understanding clinician acceptance of AI-based treatment recommendations in health care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-18). <https://doi.org/10.1145/3544548.3581075>
- Steyvers, M., & Kumar, A. (2024). Three challenges for AI-assisted decision-making. *Perspectives on Psychological Science*, 19(5), 722–734. <https://doi.org/10.1177/17456916231181102>
- Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25, 44–56. <https://doi.org/10.1038/s41591-018-0300-7>
- van Baalen, S., Boon, M., & Verhoef, P. (2021). From clinical decision support to clinical reasoning support systems. *Journal of Evaluation in Clinical Practice*, 27(3), 520–528. <https://doi.org/10.1111/jep.13541>
- van Berkel, N., Bellio, M., Skov, M. B., & Blandford, A. (2023). Measurements, algorithms, and presentations of reality: Framing interactions with AI-enabled decision support. *ACM Transactions on Computer-Human Interaction*, 30(2), 1–33. <https://doi.org/10.1145/3571815>
- van de Merwe, K., Mallam, S., & Nazir, S. (2024). Agent transparency, situation awareness, mental workload, and operator performance: A systematic literature review. *Human Factors*, 66(1), 180–208. <https://doi.org/10.1177/00187208221077804>
- Vicente, L., & Matute, H. (2023). Humans inherit artificial intelligence biases. *Scientific Reports*, 13, 15737. <https://doi.org/10.1038/s41598-023-42384-8>

- Viebahn, J., Ayedh, A., Lundberg, J., Bang, M., & Keijzers, J. (2025). User Experience Evaluation of an AI-Based Decision-Support Tool for Power Grid Congestion Management. *Human Interaction & Emerging Technologies* 2025, 197, 1. <https://doi.org/10.54941/ahfe1006694>
- Wischniewski, M., Krämer, N., & Müller, E. (2023, April). Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. In *Proceedings of the 2023 CHI conference on human factors in computing systems* (pp. 1-16). <https://doi.org/10.1145/3544548.3581197>
- Xie, Y., Chen, M., Kao, D., Gao, G., & Chen, X. A. (2020, April). CheXplain: Enabling physicians to explore and understand data-driven, AI-enabled medical imaging analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). <https://doi.org/10.1145/3313831.3376807>
- Yu, L., & Li, Y. (2022). Artificial intelligence decision-making transparency and employees' trust: The parallel multiple mediating effect of effectiveness and discomfort. *Behavioral Sciences*, 12(5), 127. <https://doi.org/10.3390/bs12050127>
- Zhang, S., Yu, J., Xu, X., Yin, C., Lu, Y., Yao, B., Tory, M., Padilla, L. M., Caterino, J., Zhang, P., & Wang, D. (2024). Rethinking human-AI collaboration in complex medical decision making: A case study in sepsis diagnosis. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2024*, 445. <https://doi.org/10.1145/3613904.3642343>
- Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020, January). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 295–305). <https://doi.org/10.1145/3351095.3372852>
- Zohrevandi, E., Westin, C. A., Vrotsou, K., & Lundberg, J. (2022, June). Exploring effects of ecological visual analytics interfaces on experts' and novices' decision-making processes: A case study in air traffic control. In *Computer Graphics Forum* (Vol. 41, No. 3, pp. 453–464). <https://doi.org/10.1111/cgf.14554>

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the publisher and/or the editor(s). This publisher and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© Copyright (2026): Author(s). The licensee is the journal publisher. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-review history:
The peer review history for this paper can be accessed here:
<https://pr.sdiarticle5.com/review-history/156846>