

Clinical explainable AI evaluations prioritize trust over effective human oversight: a scoping review

Nicolas Frey

nicolas.frey@charite.de

Charité - Universitätsmedizin Berlin <https://orcid.org/0009-0000-0195-6096>

Louis Agha-Mir-Salim

<https://orcid.org/0000-0002-2733-5084>

Linn Renner

Stefan Haufe

Lily Voge

Felix Balzer

Charité - Universitätsmedizin Berlin <https://orcid.org/0000-0003-1575-2056>

Article

Keywords: Explainable Artificial Intelligence, Clinical Decision Support Systems, Human-Centered

Posted Date: September 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10892043/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Abstract

Background Explainable artificial intelligence (XAI) is intended to help clinicians critically assess AI-supported decisions, but it is unclear whether current evaluations establish that explanations support effective human oversight. We mapped how clinician-facing XAI for clinical decision support is evaluated and whether studies assess capacities relevant to oversight.

Methods We searched PubMed, IEEE Xplore, and the ACM Digital Library for studies published from January 2015 through February 2026. Eligible studies empirically evaluated XAI in clinician-facing clinical decision support with healthcare professionals or trainees. Evaluation measures were mapped to 13 human-centered constructs and classified as objective behavioral, validated subjective, proxy or non-validated, or exploratory qualitative evidence. We additionally assessed coverage of five safety-relevant oversight capacities: objective model-behavior comprehension, appropriate reliance, failure detection, bias detection, and human–AI task performance.

Results Here we show that evaluations concentrated on user perceptions. Across 49 studies and 253 study-by-construct mappings, trust was assessed in 42 studies, human–AI task performance in 21 and appropriate reliance in 16, but failure detection in only 3, bias detection in 1 and objective model-behavior comprehension in none; 24 studies assessed none of the five capacities. Of all mappings, 138 (55%) rested on proxy or non-validated measures and 10 (4%) on validated instruments. Two studies evaluated a deployed system.

Conclusions Clinical XAI evaluation establishes how explanations are perceived far more often than whether clinicians can detect and override incorrect AI output. Demonstrating oversight requires correctness-manipulated behavioral testing, validated instruments for attitudinal constructs and evaluation that continues after deployment.

Study protocol: the protocol of this scoping review was registered on OSF on 18 December 2025 and is available at [Link].

Plain language summary

Doctors are increasingly offered AI tools that suggest a diagnosis or treatment. Because these tools can be difficult to inspect, they often add an "explanation" of how a suggestion was reached. We reviewed 49 studies that tested such explanations with doctors, nurses and other health professionals. Most studies asked how trustworthy, understandable, or useful clinicians found the explanations. Far fewer tested what clinicians actually did when the AI gave wrong advice, and none tested whether clinicians could correctly predict what the AI would do. Current evidence therefore tells us more about how explanations feel to clinicians than about whether they help clinicians catch AI mistakes. We suggest that future studies test five things directly: whether clinicians can predict model behavior, whether they accept correct advice and reject wrong advice, whether they identify individual failures, whether they recognize systematic bias, and whether clinician and AI together make better decisions.

INTRODUCTION

Artificial intelligence (AI), particularly machine learning (ML), is increasingly embedded in clinical decision support systems (CDSS), which provide clinicians with patient-specific assessments or recommendations for clinical decision-making [1, 2]. Yet the opacity of many ML models limits clinicians' ability to critically assess individual predictions. In high-stakes settings, this may promote both underreliance on useful recommendations and overreliance on erroneous, biased, or clinically inappropriate outputs [3, 4, 5]. Safe use therefore depends not only on model performance but also on clinicians' ability to exercise calibrated oversight: accepting AI advice when it is correct and rejecting it when it is wrong.

Explainable artificial intelligence (XAI) aims to support this oversight by providing information about how model outputs are generated. Methods such as Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) have advanced technical explainability [6, 7]. The regulatory relevance of such oversight is reflected in the EU AI Act, which requires high-risk AI systems to support output interpretation and effective human oversight [8].

XAI is expected to improve clinician–AI team performance by supporting more informed decision-making [9]. However, explanations are often difficult for clinicians to interpret or apply to clinical tasks, limiting their practical value [10, 11]. Technical properties such as faithfulness or completeness therefore do not automatically establish that an explanation supports safe clinical use [11].

Human-centered XAI shifts the focus from technical properties to effects on users and clinical practice, whether they improve understanding, trust, reliance, recognition of model limits, or task performance [12]. Whether this shift is realized in practice depends on which constructs are measured and how they are operationalized. Perceiving an explanation as understandable or trustworthy and being able to override an incorrect recommendation are distinct constructs. Self-reporting cannot substitute for behavioral measurement of the latter, yet it is the latter that human oversight requires [13, 14]. A clinician may report that an explanation is understandable yet still fail to reject incorrect AI advice. Without knowing which constructs current evaluations assess and what kind of evidence their measures yield, neither developers selecting explanation methods nor regulators assessing human-oversight requirements can determine what the existing evidence base supports. Existing frameworks and checklists define relevant constructs and evaluation metrics [15–17], while prior reviews summarize explanation methods, clinician perceptions, and factors influencing AI-assisted clinical decision-making [18, 19]. But the constructs and operationalizations actually used in user studies have not been mapped systematically.

This scoping review therefore mapped the methodological characteristics, clinical domains, XAI methods, evaluation constructs, operationalizations, and evidence types used in human-centered evaluations of clinician-facing CDSS. It addressed three questions: (1) What are the methodological characteristics of user studies evaluating XAI for clinician-facing CDSS? (2) Which human-centered evaluation constructs are assessed, and how are they operationalized?

(3) To what extent do current evaluations assess safety-relevant oversight capacities, including appropriate reliance, failure detection, objective model-behavior comprehension, bias or subgroup-failure recognition, and human–AI task performance?

METHODS

Study Design

The review was conducted in accordance with the JBI methodology for scoping reviews [20] and is reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) guidelines [21]. The completed PRISMA-ScR checklist is provided in Supplementary Methods (Table A1.1). The protocol was registered on OSF on 18 December 2025 and is available here: [Link].

Eligibility criteria

Eligibility criteria were defined using the Population–Concept–Context (PCC) framework and are summarized in Table 1. We included peer-reviewed primary empirical studies published from January 2015 to 22 February 2026 that involved healthcare professionals or trainees as current or prospective users of clinician-facing CDSS and empirically evaluated XAI explanations, including outcomes such as transparency, understandability, trust, satisfaction, reliance, usefulness, or human–AI task performance. We excluded studies without human-participant evaluation, studies focused exclusively on technical XAI properties, non-clinician-facing systems, non-primary studies, and grey literature, including preprints, conference abstracts without full proceedings papers, and technical reports. Grey literature was excluded to maintain a consistent publication and reporting threshold. The 2015 starting date was selected to capture contemporary XAI research, including the emergence of LIME and SHAP [6, 22].

Table 1
Eligibility and Exclusion Criteria.

Eligibility Criteria	Exclusion Criteria:
• Primary empirical study involving healthcare professionals or trainees as current or prospective users of clinician facing CDSS • Evaluation of XAI explanations in CDSS, examining attributes such as transparency, effectiveness, user trust, user satisfaction, or utility, the latter referring to the perceived or objective contribution to the decision-making process	• Focus on technical aspects of XAI without relevance to CDSS or user evaluation • Not a primary study (e.g., reviews) • Address XAI technologies designed for patients rather than healthcare professionals • Grey literature (e.g., preprints) • Presentation of prediction probabilities, confidence scores, or uncertainty estimates alone, without an explanatory component beyond the model output

Information Sources

The databases searched for this review were IEEE Xplore, ACM Digital Library, and PubMed. The initial searches were performed on 20 December 2025 and updated on 22 February 2026. Additional references were also identified through forward-backward citation search, and manual searches of key journals.

Search Strategy

The search strategy was structured according to the PCC framework (Population, Concept, Context) [23] and adapted to the syntax of each database:

- Population (P): Studies involving *health care professionals and trainees*.
- Concept (C): *XAI and CDSS*, using keywords such as “CDSS,” “Clinical Decision Support,” “Decision Support Systems,” “Artificial Intelligence,” “Machine Learning,” “Explainable AI,” “Explanations,” “Explainability,” and “Interpretability.”
- Context (C): *User evaluation and behavioral research* settings, including “User Studies,” “Vignettes,” “Survey Studies,” “Interviews,” “Trust,” “Usability Testing,” “Human–Computer Interaction,” “Cognitive Load,” and related approaches to studying human interaction with AI-based systems.

Search strings combined terms from these categories using Boolean operators (“AND,” “OR”) and were customized for each database. The search strings are available in Supplementary Methods A1.2.

Selection of sources of evidence

After duplicate removal, two reviewers independently screened titles and abstracts. Records considered potentially eligible by either reviewer underwent independent full-text assessment against the predefined criteria. Disagreements were resolved through discussion and consensus. Screening was managed in CADIMA [24].

Data Charting Process

Two reviewers independently charted data from the included studies using a predefined extraction framework. Discrepancies were resolved through discussion and consensus.

Data items

Extracted variables covered:

- Bibliographic information, including authors and publication year;
- Basic study characteristics, including study type, objectives, clinical domain, sample size, participant group, participant characteristics and their analytical use, and study context;
- AI- and XAI-related characteristics, including AI task/target type, clinician-facing output type, XAI method, explanation type, explanation scope, modality, explanation representation, and additional presentation features;
- Methodological and evaluation characteristics, including study design, evaluation approach, comparison conditions, including unaided-human and AI-without-explanation comparators, AI correctness and its manipulation, metrics used, data collection procedures, and whether the study was conducted in an experimental or real-world clinical setting; and
- The human-centered constructs assessed, their operationalizations, and the resulting operationalization categories.

Analytic framework and operationalization coding

We adapted the taxonomy from Mohseni et al. [15] to systematically characterize and map how XAI has been evaluated in clinician-facing CDSS. The framework separated the target construct from its operationalization. Each study-by-construct mapping was coded along two axes (Fig. 1): the construct axis, comprising explanation usefulness and satisfaction, understandability and mental models, trust and reliance, and human–AI task performance; and the operationalization-category axis, comprising objective behavioral, validated subjective, proxy or non-validated, and exploratory qualitative evidence. The framework derivation and cross-comparison of alternative frameworks are provided in Supplementary Methods A1.3–A1.4. Construct definitions, operationalization criteria, and coding examples are provided in Table A1.5.

The four construct dimensions were specified into 13 separately codable constructs for clinician-facing CDSS. Two substantive adaptations were made. First, reliance was divided into general reliance (TR2), defined as following, accepting, rejecting, or revising decisions toward AI advice, and appropriate reliance (TR3), defined as reliance conditioned on AI correctness and distinguishing overreliance from underreliance [13, 14, 25]. Second, model bias detection (UM4), defined as recognition of flawed, biased, or subgroup-specific model behavior, was added [26, 27]. Computational measures were excluded because they assess intrinsic properties of explanation algorithms rather than human-centered outcomes. The four operationalization categories, which we defined independently of the Mohseni et al. [15] taxonomy, were as follows

- Objective behavioral: criterion-referenced or experimentally anchored behavioral measurements, such as reliance behavior under known AI correctness, task accuracy, decision revision, response latency, failure detection, behavior under predefined model bias, or comprehension-test performance.
- Validated subjective: psychometrically established instruments, such as validated trust, usability, satisfaction, or workload scales.
- Proxy or non-validated: ad hoc Likert ratings, non-validated subjective measures, agreement rates without known AI correctness, decision changes without a criterion standard, or indirect behavioral and process proxies requiring a larger inferential step from the observed response to the target construct.
- Exploratory qualitative: open-ended interviews, think-aloud protocols, observations, focus groups, co-design sessions, and free-text feedback.

“Objective behavioral” refers to criterion-referenced or experimentally anchored performance rather than direct observability of a latent psychological state; behavioral responses without a relevant external criterion or experimental anchor were treated as proxies. The categories describe the relationship between a construct and its operationalization and were not treated as a ranked hierarchy or study-quality assessment. The classification drew on distinctions between behavioral and self-report measures of reliance [14], evidence that subjective and proxy measures can diverge from behavioral outcomes [28, 29], and considerations of construct and ecological validity. It was used to characterize evaluation practices, not to assess risk of bias.

Studies were classified as quantitative, qualitative, or mixed-methods based on their dominant analytic strategy [30, 31]. Mixed-methods classification required the collection, systematic analysis, and explicit integration of qualitative and quantitative data [32]; incidental elements of either type were insufficient. Borderline cases were resolved by consensus. Inter-rater reliability was high for construct assignment (Cohen’s $\kappa = 0.89$) and operationalization-category classification ($\kappa = 0.90$), with 91.8% raw agreement at the study-by-construct level. Remaining discrepancies were resolved through consensus.

Five evaluation targets were designated safety-relevant for clinician-facing high-risk AI systems: objective model-behavior comprehension (UM1 at the behavioral level) [33, 34], model failure detection (UM3) [35], model bias detection (UM4) [26, 27], appropriate reliance (TR3) [14, 36], and human–AI task performance (HP1) [37]. This selection was informed by the human-oversight capacities described in Article 14 of the EU AI Act[8], but represents an analytic framework for this review rather than a legally prescribed or exhaustive set of oversight requirements..

These capacities include understanding system capabilities and limitations; detecting anomalies, dysfunction, and unexpected performance; recognizing limited transferability to specific patients or subgroups; remaining aware of automation bias; and disregarding or overriding outputs while maintaining or improving task performance [8]. Comparable expectations appear in FDA, Health Canada, and MHRA guidance on good machine-learning practice and post-market monitoring [38–41].

Critical appraisal of individual sources of evidence

Consistent with the mapping objective, no formal critical appraisal or risk-of-bias assessment was conducted. Operationalization categories described how constructs were measured and were not used as study-level quality ratings.

Synthesis of Results

Study characteristics were summarized descriptively using frequencies and percentages for categorical variables and medians, interquartile ranges, and ranges for continuous variables. Human-centered evaluation was analyzed at the study-by-construct level, allowing each study to contribute multiple mappings. Construct frequencies and operationalization-category distributions were examined overall and stratified by quantitative, qualitative, and mixed-methods design. The five safety-relevant oversight capacities were additionally summarized at the study level according to how many each study assessed. No meta-analysis was performed because of substantial heterogeneity in study designs, populations, explanation methods, and outcome definitions.

Use of generative AI

Generative AI was used during manuscript preparation for language editing. The authors remain responsible for the accuracy and integrity of the final manuscript.

RESULTS

Study Selection

We identified 2,338 database records (PubMed n = 435, IEEE Xplore n = 578, ACM Digital Library n = 1,325). After removing 19 duplicates, 2,319 records were screened by title and abstract, of which 2,214 were excluded. The remaining 105 records underwent full-text assessment; 64 were excluded, leaving 41 studies identified through the database search. An additional eight eligible studies were identified through forward and backward citation searching, yielding a final corpus of 49 studies (Fig. 2). The 49 included studies are summarized in Table 2; publication frequencies by year are provided in Supplementary Results Table A2.1.

Table 2

Characteristics of the 49 included studies, ordered alphabetically by first author. All but two studies were conducted in controlled, simulated, or researcher-mediated rather than during routine clinical deployment. Study Design: Quant. = Quantitative; Qual. = Qualitative; MM = Mixed Methods.

Authors	Year	Sample Size	Clinical Domain	Study Design	XAI Method	Scope	Explanation Modality	Explanation Type	Study Object
Ackerhans et al. [47]	2025	292 Medical Students, Physicians, Nursing Professionals	Hospital care	Quant.	Not specified (generic explainability manipulation)	Local	Textual	Feature-based	Investigate explainability of AI models in hospital care, including impact on professional decision-making and patient safety.
Al-Bakri et al. [52]	2025	5 Physicians	Neurology	MM	SHAP and Grad-CAM	Local	Visual, Textual, Numerical	Feature-, Saliency-based	Evaluate the impact of explainability on clinical decision-making in Alzheimer's disease.
Anjara et al [73]	2023	10 Oncologists	Oncology	Qual.	ExampleE (example-based)	Local	Visual, Textual	Example-based	Using clinical examples to explain and improve oncology predictions.
Antoniadi et al. [53]	2022	8 Healthcare Professionals	Neurology	Quant.	SHAP	Local & Global	Visual, Textual, interactive	Feature-based	Evaluate the impact of explainability on global and local model performance, patient safety, and model usability.
Barda et al. [76]	2020	21 Critical Care Providers	Pediatric Intensive Care	Qual.	SHAP	Local	Visual, Interactive	Feature-based	Focus on the impact of explainability on model performance and patient safety in PICU.
Bergomi et al. [71]	2025	10 Clinicians	Emergency Medicine, Infectious Diseases	MM	SHAP, Araucana XAI, Bayesian Network	Local & Global	Visual, Textual, Numerical	Feature-, Rule-based, Model-inherent	Evaluate the impact of explainability on model performance and clinical decision-making.
Bienefeld et al. [42]	2023	106 Clinicians	Intensive Care	MM	SHAP	Local & Global	Visual	Feature-based	Formulate a model between development and clinical SHAP explainability.
Bussone et al. [43]	2015	8 Clinicians	Primary Care	MM	Not specified	Local	Visual, Textual	Feature-based	Investigate the impact of explainability on professional decision-making and CDSS.
Cabitz et al. [54]	2023	44 Cardiology Residents and Specialists	Cardiology (ECG interpretation)	Quant.	Local textual explanations accompanying the AI classification (not tied to a specific algorithm)	Local	Textual	Feature based	Question the impact of explainability on model performance and usability in ECG interpretation.

Authors	Year	Sample Size	Clinical Domain	Study Design	XAI Method	Scope	Explanation Modality	Explanation Type	Study Objectives
									decision response and XAI under incorrect classification
Chanda et al. [44]	2024	116 Dermatologists	Dermatology	Quant.	Grad-CAM	Local	Visual, Textual	Saliency-based	Assess domain salient enhancer confidence in melanoma diagnosis
Chen et al. [62]	2026	4 Pathologists	Pathology, Oncology	MM	Interpretable rule-set model, interactive visualization	Local & Global	Visual, Textual, interactive	Rule-, Example/Distribution-based	Develop evaluation interface explaining system's subtle reasoning
Corti et al. [77]	2024	16 Clinicians	Pulmonology	Qual.	hand-crafted explanation exemplars / conceptual XAI prototype	Local	Visual, textual	Contextual-, Workflow-based	Using exemplars to explore evolving explanations for Idiopathic Pulmonary Fibrosis
Das et al. [59]	2023	78 (16 + 62) Pulmonologists	Pulmonology	Quant.	SHAP	Local	Visual	Feature-based	Investigate collaboration impact on diagnostic accuracy and confidence in pulmonary disease
De Oliveira et al. [49]	2025	78 mental health professionals (Psychologists, Psychiatrists)	Mental Health	Quant.	LIME	Local	Visual, Textual, Numerical	Feature-, Example-based	Randomized online examination of LIME and naturalistic affect trust system detection ideation
Du et al. [69]	2022	25 Healthcare Practitioners (19 Obstetricians, 2 Midwives, 4 Dietitians)	Obstetrics (gestational diabetes)	Quant.	Feature-contribution and example-based explanations	Local	Visual, Textual, Numerical	Feature-, Example-based	Randomized within study feature contribution examination explaining gestational diabetes measurement taking of Adiponectin correction prediction and explanation preference
Famiglini et al. [56]	2024	16 Clinicians	Radiology	Quant.	Grad-CAM	Local	Visual	Saliency-based	Assess CAM's thoracic fracture in X-ray

Authors	Year	Sample Size	Clinical Domain	Study Design	XAI Method	Scope	Explanation Modality	Explanation Type	Study Objectives
Gaube et al. [74]	2023	223 Physicians	Radiology	Quant.	Visual Annotations	Local	Visual	Saliency-based	Measure effectiveness of maps in diagnosis accuracy, confidence, simulation, decision for next exper
Glick et al. [57]	2022	41 Dental Students	Dentistry	Quant.	Segmentation-based visuals	Local	Visual	Saliency-based	Evaluate assist dental radiology analy
Gombolay et al. [61]	2024	81 neurology participants (medical students, residents, and neurologists)	Neurology	Quant.	Mixed types	Local	Visual, Textual	Feature-, Rule-, Example-based	Compare different approaches to explain decisions to experts of clinical versus partic
Gomez et al. [60]	2024	121 telehealth providers: 48 physicians, 13 nurse practitioners, 60 physician assistants	Infectious Disease, Telehealth	Quant.	SHAP-like feature attribution, examples, performance metrics (mock-up prototypes)	Local + global	Visual, Textual, Numerical	Feature-, Example-based, Model performance	Evaluate different explanation forms in diagnosis accuracy under and clinical telehealth throu
Gomez et al. [63]	2025	87 Optometrists	Ophthalmology	Quant.	SHAP feature importance + intrinsically interpretable scorecard model	Local	Visual, Numerical, Textual	Feature-based (post-hoc) + Score-based (intrinsic)	Examine explanation affect performance bias to know imbalance between clinic
Guo et al. [78]	2024	2 Cardiologists	Cardiology	Qual.	Counterfactuals	Local	Visual	Counterfactual	Propose MiML cardi gener base count expla
Gutiérrez et al. [79]	2022	12 Healthcare Professionals	Nursing Homes	MM	Random Forest feature contributions + explicit rule-based explanations	Local	Visual + Textual + Interactive	Feature-based + Rule-based	Develop intera know syste priori textu; expla
Harari et al. [84]	2024	24 Emergency Medicine Clinicians	Emergency Medicine	MM	Segmented visuals	Local	Visual, Textual	Feature-based	Explore enhar (segn imagi influe and p
Hur et al. [72]	2025	63 Surgeons, Physicians with experience in prescribing blood products	Perioperative Care (blood transfusion)	Quant.	SHAP, clinical narrative explanation	Local	Visual, Textual, Numerical	Feature-based, Narrative explanation	Compare SHAP with c frienc expla no ex accep usabi transi decis

Authors	Year	Sample Size	Clinical Domain	Study Design	XAI Method	Scope	Explanation Modality	Explanation Type	Study Objectives
Ihongbe et al. [75]	2024	26 Radiologists, Other	Radiology	MM	Grad-CAM, LIME	Local	Visual	Saliency-based + Perturbation-based feature attribution	Evaluate XAI for chest diagnosis [19, p. 19]
Jabbour et al. [58]	2023	457 Clinicians	Radiology	Quant.	Grad-CAM	Local	Visual	Saliency-based	Assess whether saliency explanations reduce errors
Jacobs et al. [65]	2021	220 Clinicians	Psychiatry	Quant.	Interactive, contrastive explanations	Local	Visual, Textual	Feature-, Heuristic-based	Analyze reasons and effects of antidepressant treatment decisions
Kovalchuk et al. [80]	2022	16 Physicians	Diabetes	MM	SHAP	Local	Visual, Textual	Rule-, Feature-based, Domain reasoning	Evaluate physician acceptance of retrospective T2DM tool validation and interpretation
Kumarakulasinghe et al. [89]	2020	10 Physicians	Intensive Care	Quant.	LIME	Local	Textual	Feature-based	Compare physician reasoning with LIME explanations for sepsis prediction
Kyrimi et al. [86]	2020	16 Clinicians	Emergency Medicine	Quant.	Incremental explanation	Local & Global	Visual & Textual	Rule-based, Causal reasoning	Propose incremental (global) explanations for hybrid network clinical support
Laxar et al. [45]	2023	32 Clinicians	Critical Care	Quant.	Feature importance	Local	Visual	Feature-based	Determine whether explanations from CDSS decisions and treatment
Mandava et al. [83]	2025	n.a.	Clinical Risk Prediction	Quant.	SHAP; Partial Dependence Plots (PDP)	Local & Global	Visual	Feature-based attribution + Global feature-effect plot	Develop techniques to evaluate risk in SHAP explanations for insufficient reporting under usability trust.
Martínez-Villaseñor et al. [87]	2022	9 Clinicians	Ophthalmology	Quant.	XAHN: exact tree-based transformation of an Artificial Hydrocarbon Network with global and local feature-influence calculations	Local & Global	Visual + Numerical + Textual + Interactive	Feature-based + Rule-based / model-specific exact transformation	Develop explanations for diagnosis integrity data in global rule-based explanations
Morey and Rayo [66]	2024	30 Nurses	Emergency Medicine	Quant.	Visual annotations of predictive algorithm misalignment	Local	Visual	Feature-based	Evaluate effectiveness of misaligned display decisions in emergency scene

Authors	Year	Sample Size	Clinical Domain	Study Design	XAI Method	Scope	Explanation Modality	Explanation Type	Study Objectives
Nagendran et al. [64]	2023	86 Physicians	Intensive Care	Quant.	Feature Importance	Local	Visual, Textual	Feature-based	Compare recon in a s presc meas local impor shape doctc decis
Nagendran et al. [85]	2024	19 ICU Physicians	Intensive Care Medicine	Quant.	Q-value comparison, mortality prediction, feature importance, and training examples	Local	Visual, Textual, numerical	Contrastive/action-comparison + Outcome-based + Feature-based + Example-based	Investigates physi intera and u recon in a c decis scena sepsi CDSS tracki analy atten decis across expla
Naiseh et al. [55]	2021	16 Doctors, Pharmacists who use CDSS frequently	Prescriptions	Qual.	Not tied to a specific algorithm → interface-level explanation manipulation	Local & Global	Textual	Feature-, Example-, Counterfactual-, Confidence-based	Investigate clinic with e AI recon in CD on us leadir calibr when AI exp
Naiseh et al. [68]	2023	41 Medical Practitioners (26 Doctors, 15 Pharmacists)	Oncology (chemotherapy prescription screening)	MM	Interface-level explanation classes (local, global, example-based, counterfactual)	Local & Global	Textual, Tabular	Feature-, Example-based, Counterfactual, Global description	Within study four e class expla condi calibr agree switc behav huma perfo under incor recon follow struct interv
Nicolson et al. [50]	2025	10 Sonographers, Obstetric Registrar	Obstetrics, Sonography	Quant.	Prototype-based model (ProtoPNet)	Local	Visual	Example-based	Investigate expla a pro XAI n clinic relian diagn perfo when gesta from ultras imag
Panigutti et al. [4]	2022	28 Clinicians	Cardiology	MM	Dr. XAI	Local	Visual, Textual	Feature-based	Assess impac taking intent trust
Panigutti et al. [82]	2023	41 Clinicians	Cardiology	MM	Dr. XAI	Local	Visual, Textual	Feature-based	Refine expla interf enhar usabi

Authors	Year	Sample Size	Clinical Domain	Study Design	XAI Method	Scope	Explanation Modality	Explanation Type	Study Objective
									clinic system
Rezaeian et al. [48]	2025	28 Oncologists, Radiologists, Other	Oncology, Radiology	Quant.	Not specified (saliency-based visualization / segmentation-based explanation)	Local	Visual, Textual, Numerical	Saliency-, Example-based	Assess impact levels explain clinic report diagn accuracy the effectiveness of clinic decision making
Sabol et al. [46]	2020	14 Pathologists	Pathology, Oncology	Quant.	Cumulative Fuzzy Class Membership Criterion (CFCMC)	Local	Visual, Textual	Example-, Possibility-based	Evaluate explainability of class color histology using retrospection and patient feedback
Schwartz et al. [81]	2022	17 Clinicians	Hospital Management	Qual.	Feature based visual explanation	Local & Global	Visual	Feature-based	Explore trust prediction using baseline in-hospital data
Singh et al. [70]	2021	14 Clinicians	Ophthalmology	Quant.	Attribution methods (Deep Taylor, Guided Backpropagation, SHAP, IG, LRP, Occlusion, etc.)	Local	Visual	Saliency-, Feature-based	Evaluate comparison multiple explainability for re-diagnosis assess clinician mean and a
Sivaraman et al. [67]	2023	24 ICU Clinicians	Critical Care	MM	SHAP + Alternative Treatments visualization	Local	Visual, Textual	Feature-, Example-based	Investigate ICU clinician interaction generate treatment recommendations how it affects and a behavior pattern human decision
Wysocki et al. [51]	2023	23 Healthcare Professionals	Oncology	Quant.	SHAP	Local	Visual	Feature-based	Evaluate impact explain machine output health professional trust, understanding and decision making (COR hospital admission decision with C
Zou et al. [88]	2023	3 Radiologists	Radiology	MM	SHAP, Grad-CAM	Local & Global	Visual	Saliency-, Feature-based	Evaluate clinic local effectiveness an explainability algorithm comb

Authors	Year	Sample Size	Clinical Domain	Study Design	XAI Method	Scope	Explanation Modality	Explanation Type	Study Object
									and COVID-19 X-ray

Methodological characteristics of the included studies

Study objectives

Across the 49 included studies, five recurring non-mutually exclusive research objectives could be distinguished. The most frequent objective concerned the impact of XAI on clinicians' trust and perception of CDSS (N = 15), investigating whether and how the presence, format, or complexity of explanations influences clinicians' subjective evaluation, including perceived reliability, confidence, or acceptance [4, 42–55].

Two objectives were equally frequent. Clinical decision-making and task performance was examined in N = 14 studies, assessing whether XAI affects diagnostic accuracy, treatment decisions, or overall task outcomes when clinicians interact with AI-based CDSS [43, 48, 50, 51, 54, 56–64].

Reliance behavior and decision calibration was likewise examined in N = 14 studies, measuring not only whether clinicians followed AI advice but whether they did so appropriately, distinguishing overreliance on incorrect advice from underreliance on correct recommendations [43, 44, 50, 51, 54, 55, 58, 63–69].

The next most frequent objective concerned comparison of explanation formats or methods (N = 13), evaluating which types of explanations clinicians found most useful, understandable, or actionable [48, 49, 55, 56, 61, 68–75]. The fifth objective concerned system development, co-design, or clinician needs (N = 11), prioritizing identification of explanation requirements and design preferences over summative outcome evaluation [4, 42, 55, 73, 76–82].

Study designs, sample sizes, and participants

Twenty-nine studies used quantitative methods, six qualitative, and 14 mixed-methods designs. One study [83] did not report participant numbers and was excluded from all participant-level summaries (effective n = 48). Across the remaining studies, participant numbers ranged from 2 to 457 (median = 24, IQR = 13.5–66.8; Table 2). Quantitative studies included 8–457 participants (median = 36.5), qualitative studies 2–21 (median = 16), and mixed-methods studies 3–106 (median = 20).

Participants were predominantly licensed clinicians and healthcare professionals. Two studies primarily involved trainees: Glick et al [57] included 41 dental students exclusively, whereas Ackerhans et al. [47] included predominantly medical students (278 of 292 participants) alongside physicians and nursing professionals. Other participant groups included nurses [66], pharmacists [55, 68], telehealth providers [60], optometrists [63], pathologists [46, 62], psychologists or psychiatrists [49] and obstetricians, midwives, and dietitians [69]. Participant characteristics were frequently reported but less consistently incorporated analytically. Clinical seniority, experience, or training level was analytically incorporated in 20 studies [42, 44, 45, 48–51, 54, 56, 57, 59–61, 64, 65, 67, 69, 71, 72, 74]. Professional role or specialty was incorporated into analyses in 11 studies [42, 49, 58, 60, 65, 67, 69, 71, 72, 74, 76].

AI-related background was characterized in 14 studies [48–50, 55, 58, 61, 65, 68, 69, 71, 72, 74, 76, 84] and analytically incorporated in eight [48–50, 58, 61, 65, 69, 76], including predictive-modeling knowledge, AI/XAI familiarity, objective or self-reported AI knowledge, and prior CDSS use. Only De Oliveira et al. [49] included an objective AI-knowledge test, and no study reported use of a named validated AI-literacy instrument.

Clinical domains

Critical care and intensive care medicine were most frequently represented (n = 7), followed by radiology (n = 6), oncology and pathology (n = 6), emergency medicine (n = 4), cardiology (n = 4), and ophthalmology and neurology (n = 3 each). Other domains included dermatology, dentistry, pulmonology, psychiatry, diabetes, primary care, nursing homes, hospital management, obstetrics, blood transfusion decision-making, mental health, telehealth, and infectious diseases (Table 2).

Study setting

Two study setting dimensions distinguished the included studies: the degree of experimental control over AI behavior, and the proximity to clinical practice. Forty-seven of 49 studies (96%) evaluated XAI in controlled, simulated, or researcher-mediated settings rather than during routine clinical deployment. These included vignette-based surveys, reader studies, scenario-based simulations, case-evaluation tasks, and qualitative or co-design evaluations. Twenty of 49 studies exposed participants to both correct and incorrect AI outputs: 15 quantitative [44, 48, 50, 51, 54, 56–59, 61, 63, 65, 66, 69, 85], 4 mixed-methods [43, 62, 68, 71], and 1 qualitative study [55]. In eight of these studies, AI correctness itself was explicitly manipulated as an experimental factor, enabling a more direct causal assessment of appropriate versus inappropriate reliance (Supplementary Results Table A2.5).

Two of 49 studies (4%) evaluated deployed CDSS in their target clinical environment: Schwartz et al. observed clinician interaction with a deterioration prediction system in a hospital ward [81], and Gutiérrez et al. instrumented a recommendation system in nursing homes over a four-month deployment [79]. No study longitudinally assessed how explanation-related outcomes or clinician interaction with explanations changed over repeated exposure in routine clinical use.

Seventeen studies objectively measured how accurately clinicians performed when using AI [43, 44, 48, 50, 51, 56–63, 65, 68, 74, 84]. However, only nine also measured clinician performance without AI [44, 48, 50, 57–59, 62, 63, 65], allowing direct comparison of performance with and without AI support. The remaining eight [43, 51, 56, 60, 61, 68, 74, 84] did not include a clinician-only condition, so they could not directly compare AI-assisted performance with

unaided clinician performance. Separately, 23 of 49 studies provided an AI-without-explanation comparison alongside an explainable condition. In 15 studies, the explanation was the principal manipulated difference; six used sequential or fixed exposure in which the AI output was shown before the explanation or explanation levels were introduced in a fixed order, and two changed the underlying model as well as explainability. Among the 17 studies with objective criterion-based human–AI task performance, 11 included an AI-without-explanation comparator; six were direct comparisons, four were sequential or fixed-order comparisons, and one was model-confounded.

Explanation scope, methods, and modalities

Thirty-seven studies evaluated local explanations only, while 12 combined local and global components [42, 53, 55, 60, 62, 68, 71, 81, 83, 86–88], and none evaluated global explanations alone (Table 2). Feature-based explanations dominated, with SHAP as the single most frequent technique (Fig. 3). SHAP was used either alone [42, 51, 53, 59, 76], or layered with other approaches, such as Grad-CAM [52, 88], clinical narratives [72], scorecards [63], hybrid rule-based and probabilistic methods [71], alternative-treatment visualizations [67], partial dependence plots [83] or a broader set of compared attribution methods [70]. Saliency methods, particularly Grad-CAM, clustered in imaging studies [44, 52, 56, 58, 70, 75, 88], alongside segmentation overlays [57], visual annotations [74], and augmented-reality overlays [84]. Rule-based, counterfactual [68], and example based [68, 69] or prototype-based explanations were less common. Several studies layered explanation types. Naiseh et al. [55], for instance combined feature-, example-, counterfactual-, and confidence-based explanations with a global model description. A complete overview of the evaluated XAI methods, grouped by method family, is provided in Supplementary Results Table A2.4.

Presentation modality was predominantly visual or visual–textual; purely textual explanations were used in 5 studies. Five studies included interactivity [53, 62, 76, 79, 87], one used augmented reality overlays [84], and ten additionally included numerical elements such as probabilities, confidence values, or risk decompositions [48, 49, 52, 60, 63, 69, 71, 72, 85, 87].

Tabular feature-importance displays (bar plots, tornado plots, top-k lists, SHAP contribution plots) dominated structured-data tasks [45, 51, 53, 59, 69, 76]; pixel-level overlays (heatmaps, CAMs, segmentation masks, prototype patches) dominated imaging tasks [44, 50, 56–58, 70]; and structured symbolic forms (IF–THEN rule sets, decision trees, causal/Bayesian graphs) carried the rule- and reasoning-based explanations [4, 62, 71, 86, 87].

Human-Centered Evaluation Practices of XAI

To characterize how XAI for CDSS was evaluated, each included study was mapped to the 13-construct framework. Because a study could assess several constructs, the 49 studies yielded 253 construct mappings. Studies assessed a median of five constructs (IQR 4–6; range 2–9), with similar coverage across quantitative (median 5), qualitative (median 6), and mixed-methods studies (median 6). Each mapping was additionally classified by operationalization category. We report results in three steps: which constructs were evaluated, how they were operationalized and classified by study design, and how the corpus covered the five safety-relevant oversight capacities. Full study-level construct and operationalization assignments underlying these aggregate results are provided in Supplementary Methods Table A1.6.

Construct Coverage

Coverage was highly uneven across the 13 constructs (Fig. 4). Trust was the most frequently assessed construct (42 of 49 studies; 23 of 29 quantitative studies), followed by explanation usefulness (34 studies) and perceived understandability (33 studies). These three co-occurred in 26 of 49 studies and together accounted for 109 of the 253 mappings (43%). Comparison or preference between explanation formats was assessed in 27 studies, general reliance in 29, and explanation satisfaction in 21.

Other constructs were less frequently represented. Perceived completeness of explanations was assessed in 11 studies, task-completion time (HP2) in 8, and perceived cognitive load in 7. Overall, the distribution indicates that evaluations concentrated primarily on trust, perceived usefulness, understandability, and general interaction with AI advice, whereas several constructs concerned with model limitations and correctness-sensitive behavior were much less frequently examined. The subset of constructs defined as safety-relevant is summarized separately below.

Operationalization and Evidence type

Across the 253 mappings, 120 (47%) used subjective operationalizations, 88 (35%) used objective behavioral or process measures, and 45 (18%) drew on exploratory qualitative evidence. Among the 208 mappings based on structured quantitative measures, 120 (58%) were subjective and 88 (42%) objective or behavioral. Of the 120 subjective mappings, 10 used validated instruments and 110 relied on non-validated self-report. Of the 88 objective or behavioral mappings, 60 directly operationalized the target construct and 28 represented indirect behavioral or process proxies. Overall, proxy or non-validated evidence was most common (138/253, 55%), followed by objective behavioral evidence (60/253, 24%), exploratory qualitative evidence (45/253, 18%), and validated subjective evidence (10/253, 4%). This pattern was similar across quantitative and mixed-methods studies; all mappings from qualitative studies were classified as exploratory qualitative evidence (Fig. 5).

Objective behavioral measurement designs were used mainly for task performance, task-completion time, and reliance, whereas trust, explanation usefulness, and perceived understandability were measured predominantly with proxy or non-validated self-report. Validated measures were sparse (Fig. 5). We next describe the four evidence types before summarizing the safety-relevant oversight capacities.

Objective behavioral measures

Objective behavioral evidence accounted for 60 of 253 mappings (24%). Human–AI task performance, usually scored as accuracy against ground-truth labels, was the most common direct behavioral outcome and was objectively assessed in 17 studies. Reliance was captured behaviorally in seven studies

that used Weight-of-Advice (WoA) paradigms to quantify how strongly participants revised an initial unaided decision after seeing the AI output [4, 45, 50, 69, 72, 82, 85]. As a direct measure of decision revision, WoA was coded as an objective behavioral operationalization of general reliance (TR2) in all seven. It supported appropriate reliance (TR3) in three studies where revision was conditioned on AI correctness [50, 69, 85]; in the remaining four, WoA was aggregated irrespective of correctness. Five further studies operationalized appropriate reliance directly without WoA, using correctness-sensitive agreement, switching, or decision-revision measures [43, 54, 59, 61, 68].

Of the 20 studies that exposed participants to both correct and incorrect AI outputs, 16 assessed appropriate reliance: eight used direct objective behavioral operationalizations, seven used proxy or non-validated measures, and one used exploratory qualitative evidence. The remaining four studies did not assess appropriate reliance as a distinct construct. Sixteen of the 20 also measured human–AI task performance. Efficiency was operationalized using task-completion time or response latency in eight studies; one study additionally incorporated eye-tracking [85]. Comparison between explanation formats was operationalized behaviorally in 14 studies [43, 47–49, 56, 60, 61, 64, 65, 67–69, 72, 84], in which explanation conditions were experimentally manipulated and contrasted rather than rated for preference.

Behavioral measures of model-limit recognition were rare. No study included an objective model-behavior comprehension test, such as a forward-simulation or output-prediction task. Failure detection was represented by two proxy measures [66, 68] and qualitative assessment [80]. Bias detection was represented by one objective behavioral assessment using a manipulated biased-versus-correct AI design and diagnostic performance scored against ground truth [58].

Validated subjective instruments

Validated subjective evidence accounted for 10 of 253 mappings (4%) across seven studies; several instruments contributed mappings to more than one construct (Supplementary Results Table A2.2). Trust-related mappings were operationalized using the *Trust in Automation Scale* [84], the *Trust Scale Recommended for XAI* [72], an adapted version of a validated trust scale [47], and the *Human–Computer Trust Scale* [68], the latter also contributing a validated understandability mapping. Explanation satisfaction was operationalized with the *Explanation Satisfaction Scale* [4, 72, 82], one study jointly with the *System Usability Scale* [72] and the *System Causability Scale* provided validated understandability and completeness mappings [49]. In Panigutti et al. [4, 82], construct-specific usefulness, understandability, and completeness mappings derived from individual *Explanation Satisfaction Scale* items were classified as proxy or non-validated rather than validated subjective evidence. In De Oliveira et al. [49], adapted *Human–Computer Trust Model* items were used to assess trust but were classified as proxy or non-validated rather than validated subjective evidence. The *Negative Attitudes toward Robots Scale* was used by Gombolay et al. [61], but was not counted as a validated trust operationalization because it measures negative attitudes toward robots rather than trust in the target construct. Other established questionnaires used in the corpus measured constructs outside the review framework, including the *Godspeed Questionnaire* [84]. Three further studies drew on the *UTAUT* model (Unified Theory of Acceptance and Use of Technology) [4, 42, 82], but used it to capture technology acceptance and intention to use rather than as a validated instrument for a construct included in the framework.

Proxy measures and non-validated measures

Proxy or non-validated evidence accounted for 138 of 253 mappings (55%), making it the most common operationalization category. Attitudinal and perceptual constructs were typically rated on ad-hoc 5- to 10-point Likert scales, with two studies using semantic differentials [46, 87], applied above all to trust, satisfaction, explanation usefulness, perceived understandability, and format preference. Reliance was often measured only as raw agreement or concordance with AI recommendations [71, 80, 89], as diagnostic revisions [60], prescription shifts, decision switches [67], or behavioral uptake without varying AI correctness [62]. These measures were classified as proxy operationalizations under the framework.

Explorative qualitative evidence

Explorative qualitative evidence accounted for 45 of 253 mappings (18%). Qualitative operationalizations drew on think-aloud protocols [43, 55, 67, 73], semi-structured interviews [42, 55, 77, 80, 81], focus groups [42, 76], co-design and group-review sessions [78, 82], open-ended survey responses [53, 70, 75], and observational notes [55]. These approaches primarily addressed satisfaction, usefulness, and perceived understandability.

A complete inventory of behavioral and objective measures is given in Supplementary Results Table A2.3, and the full list of validated instruments in Supplementary Results Table A2.2.

Safety-relevant oversight capacities

Among the five safety-relevant oversight capacities, human–AI task performance was assessed most frequently in 21 of 49 studies, followed by appropriate reliance under varying AI correctness in 16, failure detection in 3 [66, 68, 80], and bias detection in a single study using a manipulated biased-versus-correct design [58]; objective model-behavior comprehension, for example forward-simulation or output-prediction tasks, was assessed in none. Of the 49 studies, 24 assessed none of these capacities, 22 assessed one or two, and 3 assessed three; none assessed four or all five capacities. Direct objective evidence was available for 17 task-performance mappings, eight appropriate-reliance mappings, and the single bias-detection mapping, but for neither failure detection nor objective model-behavior comprehension.

DISCUSSION

Principal Findings

We examined how human-centered XAI for clinician-facing CDSS is evaluated across 49 studies published between January 2015 and February 2026. Clinical human-centered XAI evaluation showed more clearly how clinicians perceived explanations than whether explanations supported effective oversight. Subjective measures dominated structured evaluations, while human–AI task performance and appropriate reliance were assessed in fewer than half of the studies and with direct behavioral measures in fewer still, whereas failure detection, bias detection, and objective model-behavior comprehension were

assessed in three studies, one study, and none, respectively. Subjective constructs were usually measured with ad hoc or proxy measures rather than validated instruments. Together, these findings reveal a persistent gap between safety-relevant claims about XAI and the measures used to substantiate them.

Methodological Landscape

Almost all studies were controlled or simulated; only two evaluated deployed CDSS [79, 81]. These designs improve experimental control but limit ecological validity: single encounters cannot establish whether explanation effects persist under time pressure, competing priorities, repeated exposure, or automation fatigue, the conditions under which oversight may be most vulnerable [90, 91]. Comparison conditions were a second gap. Many studies lacked either an unaided-human baseline or a clean AI-without-explanation comparator, and some comparisons were sequential, fixed-order, or model-confounded. This limited the ability to separate the benefit of AI support from the added benefit of the explanation.

Samples were also often small, single-site, and specialist-heavy, with limited interprofessional representation, echoing Abbas et al. [92], leaving settings such as primary care and nursing comparatively underexamined. Because expertise can moderate AI-supported decision behavior [74], reliance on small, specialist-heavy samples may therefore obscure clinically relevant heterogeneity in how novices, generalists, and non-specialists respond to the same explanations. Participant characteristics were usually reported but less often analyzed, most commonly seniority or training level, followed by professional role or specialty and AI-related background; analyses were predominantly descriptive or covariate-based rather than explicit moderator models. AI-related background was also almost always self-reported, and AI competencies are not yet systematically established in medical training [93], so familiarity cannot be assumed uniform within a clinician sample.

More broadly, heterogeneity and incomplete reporting of contextual factors such as interface design, case difficulty, workload, and perceived system authority often limited interpretation of explanation-specific effects. Explanation methods were also concentrated. Feature-attribution approaches, particularly SHAP, LIME, and saliency methods such as Grad-CAM, predominated, whereas counterfactual and example-based approaches were less frequent and adaptive explanations were absent. This matters because the most frequently evaluated attribution and saliency methods are also among those whose technical validity has been contested [94–98] (see Critical perspective on the XAI premise below), while evidence on whether richer interactive formats better support clinical reasoning remains limited.

Why perceptual measures cannot establish safety

Trust remains an important human-factors outcome because insufficient trust can contribute to underreliance on correct AI advice. However, trust ratings alone cannot establish whether a system is used safely, because trust reflects an attitude toward the system whereas reliance is the behavioral act of accepting, revising, or overriding its advice [13, 14]. The same level of reported trust may therefore accompany appropriate reliance on correct advice or overreliance on incorrect advice. Moreover, reported trust does not reliably predict observed reliance [28, 29]. One recent study found no relationship between trust and reliance [99], consistent with broader evidence that self-reported and observed behavior diverge [100]. Trust-based comparisons may therefore favor explanations that feel convincing over those that would support safer decision-making.

Appropriate reliance was assessed less often than trust (16/49 vs. 42/49 studies), and only eight studies used direct behavioral measures sensitive to AI correctness. Reliance becomes safety-relevant only when AI correctness is known or manipulated: higher agreement can reflect either appropriate uptake of correct advice or overreliance on wrong advice [101]. Measures such as Relative AI Reliance and Relative Self-Reliance distinguish acceptance of correct from rejection of incorrect advice [101]. Even among studies that exposed participants to both correct and incorrect AI outputs, appropriate reliance was assessed inconsistently, and only about half used direct objective measures.

Understandability showed an even sharper mismatch. Although frequently evaluated, no study used a purpose-built model-behavior comprehension task; assessments instead relied on subjective ratings or qualitative evidence. Failure detection was similarly rarely assessed with direct behavioral measures, while bias detection was examined in only one study using a correctness-manipulated behavioral design [58]. This matters because explanations may increase perceived understanding without improving actual comprehension or error detection [102]. Criterion-referenced tasks such as forward simulation or output prediction can show whether users anticipate model behavior, while experimentally introduced failures or biases can test whether they detect, reject, or resist problematic outputs. However, such tasks cannot directly establish whether users understand why the model behaves as it does. Jabbour et al. [58], for example, objectively measured behavior under manipulated model bias without establishing explicit bias recognition; likewise, forward simulation measures anticipation of model behavior rather than causal or clinical understanding [33, 34]. From a safety perspective, however, explicit understanding may be less important than whether users reliably translate available information into appropriate decisions. Behavioral success should therefore be treated as evidence of effective oversight, not necessarily of explicit model understanding. The same applies beyond explanations. Model-level information is not uniformly understood by those who receive it [103], and it is not inert: disclosing characteristics of the training data can change how physicians respond to model output [104]. Whether such information is understood, and what it does to behavior, are separate empirical questions.

Buçinca et al. [28] showed that proxy-task performance did not predict actual decision-task performance and that subjective trust and preference ratings likewise failed to predict objective performance. The problem is not that subjective measures are inherently inappropriate: they are appropriate for constructs such as trust or satisfaction, but safety-relevant oversight capacities require behavioral evidence. Qualitative evidence can identify mechanisms and failure modes but cannot by itself establish behavioral performance; conversely, legitimately subjective constructs should ideally be assessed with validated instruments.

The ability of explanations to reduce overreliance may depend on whether verifying the AI output is less effortful than independently solving the task [105], a condition that may often be unfavorable in clinical settings. Perceived cognitive load was rated in seven studies, but never with a validated workload instrument such as NASA-TLX [106], and completion time was treated as efficiency rather than as verification effort.

Critical perspective on the XAI premise

Evaluating XAI requires clarity about the functions that explanations are intended to serve. The reviewed studies reported beneficial, absent, and adverse behavioral effects that varied across tasks and clinicians [50, 60, 74]. Thus, explanations cannot be assumed to improve understanding, calibrated trust, oversight, reliance, or safety. Their presence may instead provide reassurance or signal competence and fairness without enabling verification [107], particularly when checking an explanation is as demanding as independently solving the clinical task [36, 105].

The value of post-hoc explanations is also contested. Some argue that rigorous clinical validation addresses opacity more effectively than explanations that only approximate model reasoning [94], whereas others consider such approximations useful when properly evaluated [108]. Attribution methods may highlight features unrelated to the prediction, and Shapley-based methods can rank irrelevant features above those actually driving it [95–97]. Saliency methods likewise may fail sanity checks or align poorly with ground-truth abnormalities [98, 109]. Conversely, even a faithful explanation may highlight clinically irrelevant features; in some cases, such features can reflect shortcut learning or other non-clinical statistical dependencies [110, 111], but they may also serve suppressor, normalization, or robustness-enhancing roles [112]. Attribution alone therefore does not establish the statistical function of a highlighted feature or whether its use constitutes a model failure. Faithfulness itself is not captured by a single accepted formal criterion [113], and empirical evaluations can be influenced by properties of the input data: recent work showed that attribution methods may emphasize visually salient features irrespective of their task informativeness [114]. XAI may therefore generate hypotheses about model failures, but attribution alone cannot establish that a failure has occurred [94, 95, 97, 109].

Moreover, users may correctly learn which features a model responds to while drawing incorrect clinical or causal inferences from these patterns, potentially reinforcing rather than correcting pre-existing beliefs [1]. Concept-based methods such as Testing with Concept Activation Vectors (TCAV) likewise do not by themselves establish relevance, causality, or actionability [115].

These technical and human-centered problems mirror each other. Human-centered evaluation may reward explanations that clinicians find understandable, while technical evaluation may reward explanations that appear faithful despite lacking clinical meaning. Neither standard alone establishes safer decisions when CDSS advice is wrong or misleading. Behavioral evaluation cannot repair an unfaithful explanation, but it can determine whether clinicians can nevertheless exercise effective oversight.

Implications for evaluation practice

Evaluation measures should be selected to match the construct an explanation is intended to affect. Despite broader claims about understanding, reliance, and oversight, evaluations largely focus on perceived usefulness, trust, and understandability. Future evaluations should align XAI claims with the constructs and measures used and limit conclusions accordingly.

Existing tools address complementary aspects of evaluation. CLIX-M provides metrics for properties of clinical explanations [17], whereas DECIDE-AI addresses human factors and safety in early clinical evaluation [116]. Our proposal is narrower: the five capacities examined here form a candidate set of proximal oversight capacities, derived by linking externally defined human-oversight functions [8] to constructs and measurement paradigms from prior human–AI research [13, 26, 33, 35–37]. Illustrative behavioral operationalizations and interpretation boundaries for each capacity are summarized in Table 3. They are neither a validated checklist nor a universally sufficient safety standard. Their relative importance and operationalization will depend on system risk, evaluation purpose, and development stage. Reporting which capacities were assessed—and justifying relevant omissions—could improve alignment between claims, empirical evaluation, and regulatory expectations.

Table 3
Candidate operationalizations for five safety-relevant human-oversight capacities. Examples are illustrative rather than prescriptive and should be adapted to the system, task, risk, and stage of evaluation.

Oversight capacity	Illustrative behavioral operationalization	Interpretation boundary
Model-behavior comprehension (UM1)	Forward simulation or output prediction on new cases; score the predicted AI output against the model’s actual output.	Shows whether users can anticipate model behavior; does not establish causal or clinically valid understanding.
Appropriate reliance (TR3)	Vary correct and incorrect AI advice and separately quantify acceptance of correct advice and rejection of incorrect advice (e.g., correctness-conditioned WoA, RAIR/RSR).	Requires known AI correctness; aggregate agreement or uptake alone cannot distinguish appropriate reliance from overreliance.
Failure detection (UM3)	Present predefined erroneous or unreliable AI outputs and assess explicit detection or correctness-sensitive rejection; score detection performance where possible.	Explicit flagging provides the strongest evidence that the failure itself was recognized; correctness-sensitive rejection also provides objective behavioral evidence of successful oversight.
Bias / subgroup-failure detection (UM4)	Introduce a predefined systematic subgroup-specific error pattern and assess whether users identify or behaviorally resist affected outputs; Jabbour et al. [58] illustrates the latter.	Behavioral resistance is objective evidence of the response to biased AI; explicit identification provides stronger evidence that the bias itself was recognized.
Human–AI task performance (HP1)	Score clinical task performance against a criterion; ideally compare unaided human, AI-only, and AI + XAI conditions.	Shows task outcome, not the mechanism; appropriate comparators are needed to isolate the contributions of AI and explanation.

How these capacities should be operationalized depends on the stage of evaluation. Before deployment, correctness-manipulated experiments can assess appropriate reliance and failure detection, subgroup-specific errors can test bias detection, and forward-simulation tasks can assess model-behavior

comprehension. Correctness-sensitive probes should remain non-actionable, for example in simulation or retrospective review. After deployment, longitudinal evaluation should examine reliance, workflow integration, and human–AI task performance under routine conditions, supported by process data such as decision- and interaction logs and response times. No study in the corpus used such a multimethod longitudinal design. Decision-theoretic measures could further evaluate reliance against payoff-based baselines and rational benchmarks rather than simple accept–reject counts [117].

Explanations are also not the only transparency mechanism available to clinicians. Model-level documentation [118] and communicated predictive uncertainty [119, 120] address capabilities, limitations, and confidence that instance-level attribution does not. Future work should evaluate these mechanisms in combination and against the same oversight capacities.

Limitations

This review has several limitations. The search covered three databases (IEEE Xplore, ACM Digital Library, and PubMed) from January 2015 until February 2026 and excluded grey literature; relevant preprints or conference abstracts without full text may have been missed, so the true frequency of behavioral or validated measures could be somewhat higher than reported here if newer, unpublished evaluation approaches exist outside this search window. As a scoping review, its aim is to map the evidence rather than perform a risk-of-bias assessment or an effect-size synthesis, so this review cannot say whether methodologically stronger studies showed a different balance of measures than weaker ones; the reported pattern is descriptive of the corpus as a whole, not adjusted for study quality. The framework for the construct mapping also involved interpretive choices: although the Mohseni et al. [15] taxonomy was selected for its breadth and uptake, other frameworks could weight or group constructs differently, so the exact proportions reported here, particularly the boundary between objective behavioral and proxy operationalizations, are to some extent framework-dependent. The boundary of the safety-relevant set is likewise interpretive: a narrower definition excluding model-behavior comprehension, on the grounds that it is a precondition for oversight rather than an oversight outcome, would also be defensible, and would not change the observed pattern. A further boundary concerns what "safety-relevant" can mean here. The constructs we identify are proximal: they concern clinician behavior toward AI output, not patient outcomes. No study in this corpus measured harm, and the inferential chain from miscalibrated reliance to clinical injury remains untested empirically. Our claim is therefore that current evaluation does not establish the preconditions for safe oversight, not that it demonstrates harm. A further limitation concerns construct overlap. Some capacities are conceptually adjacent, particularly failure detection (UM3) and appropriate reliance under incorrect advice (TR3), both of which involve clinicians' responses to erroneous predictions. Assignment of individual measures therefore required interpretive judgment and may have affected counts at the margins.

CONCLUSION

Across 49 studies, human-centered evaluation of XAI for clinical decision support concentrated on attitudinal and perceptual outcomes, particularly trust, perceived usefulness, and understandability. Among structured quantitative measures, subjective operationalizations modestly predominated, while several safety-relevant oversight capacities were rarely or never assessed with direct behavioral measures. The literature therefore shows more clearly how clinicians perceive explanations than whether explanations improve decisions when AI output is wrong. Closing this gap requires behavioral measurement under manipulated AI correctness, validated instruments where constructs are legitimately attitudinal, and evaluation that continues after deployment. Without this shift, evaluations may continue to demonstrate favorable user perceptions while providing limited evidence about whether explanations support effective human oversight under clinically consequential model errors.

Declarations

Ethics approval and consent to participate

Not applicable. This scoping review synthesized data from published studies and did not involve the recruitment of human participants or the collection of identifiable personal data.

Consent for publication

Not applicable.

Data availability

All data underlying the review findings are derived from the cited publications. Study-level coding assignments and supplementary analyses are provided in the Supplementary Information.

Competing interests

The authors declare no competing interests.

Funding

N.F., S.H., L.A.M.S., L.R., L.I.M.V. and F.B. declare no relevant funding.

Authors' contributions

N.F. conceived and designed the review, developed the methodology and search strategy, conducted the literature search, screening, data extraction and coding, performed the analyses and synthesis, developed the figures and tables, and drafted the manuscript. L.R. supported study identification, data

extraction, and coding. L.I.M.V. supported study identification and data extraction. L.A.M.S supported study identification, data extraction, and coding. S.H. and F.B. critically reviewed and revised the manuscript. All authors reviewed and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Abell B, Naicker S, Rodwell D, Donovan T, Tariq A, Baysari M, et al. Identifying barriers and facilitators to successful implementation of computerized clinical decision support systems in hospitals: a NASSS framework-informed scoping review. *Implement Sci.* 2023;18(1):32. doi:10.1186/s13012-023-01287-y
2. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. *Npj Digit Med.* 2020;3(1):17. doi:10.1038/s41746-020-0221-y
3. Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lerner E, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *Npj Digit Med.* 2021;4(1):31. doi:10.1038/s41746-021-00385-9
4. Panigutti C, Beretta A, Giannotti F, Pedreschi D. Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems. In: *CHI Conference on Human Factors in Computing Systems [Internet]*. New Orleans LA USA: ACM; 2022 [cited 2023 Oct 1]. p. 1–9. Available from: <https://dl.acm.org/doi/10.1145/3491102.3502104> doi:10.1145/3491102.3502104
5. Dietvorst BJ, Simmons JP, Massey C. Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Manag Sci.* 2018;64(3):1155–70. doi:10.1287/mnsc.2016.2643
6. Ribeiro MT, Singh S, Guestrin C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]*. San Francisco California USA: ACM; 2016 [cited 2025 Jan 10]. p. 1135–44. Available from: <https://dl.acm.org/doi/10.1145/2939672.2939778> doi:10.1145/2939672.2939778
7. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions.
8. European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). 2024.
9. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion.* 2020;58:82–115. doi:10.1016/j.inffus.2019.12.012
10. Abgrall G, Holder AL, Chelly Dagdia Z, Zeitouni K, Monnet X. Should AI models be explainable to clinicians? *Crit Care.* 2024;28(1):301. doi:10.1186/s13054-024-05005-y
11. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health.* 2021;3(11):e745–50. doi:10.1016/S2589-7500(21)00208-9
12. Schoonderwoerd TAJ, Jorritsma W, Neerincx MA, Van Den Bosch K. Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. *Int J Hum-Comput Stud.* 2021;154:102684. doi:10.1016/j.ijhcs.2021.102684
13. Lee JD, See KA. Trust in Automation: Designing for Appropriate Reliance. *Hum Factors.* 2004.
14. Vereschak O, Bailly G, Caramiaux B. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proc ACM Hum-Comput Interact.* 2021;5(CSCW2):1–39. doi:10.1145/3476068
15. Mohseni S, Zarei N, Ragan ED. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Trans Interact Intell Syst.* 2021;11(3–4):1–45. doi:10.1145/3387166
16. Hoffman RR, Mueller ST, Klein G, Litman J. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Front Comput Sci.* 2023;5:1096257. doi:10.3389/fcomp.2023.1096257
17. Brankovic A, Cook D, Rahman J, Delaforce A, Li J, Magrabi F, et al. Clinician-informed XAI evaluation checklist with metrics (CLIX-M) for AI-powered clinical decision support systems. *Npj Digit Med.* 2025;8(1):364. doi:10.1038/s41746-025-01764-2
18. Jackson NJ, Brown KE, Miller R, Murrow M, Cauley MR, Collins BX, et al. Factors influencing the effectiveness of artificial intelligence-assisted decision-making in medicine: a scoping review. *J Am Med Inform Assoc.* 2026;33(5):1054–64. doi:10.1093/jamia/ocag002
19. Antoniadis AM, Du Y, Guendouz Y, Wei L, Mazo C, Becker BA, et al. Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Appl Sci.* 2021;11(11):5088. doi:10.3390/app11115088
20. Peters MD, Godfrey C, McInerney P, Munn Z, Tricco AC, Khalil H. Scoping reviews. In: Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z, editors. *JBIM Manual for Evidence Synthesis [Internet]*. JBI; 2024 [cited 2026 Jul 20]. Available from: <https://jbi-global-wiki.refined.site/space/MANUAL/355862497/10.+Scoping+reviews> doi:10.46658/JBIMES-24-09
21. Tricco AC, Lillie E, Zarin W, O’Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann Intern Med.* 2018;169(7):467–73. doi:10.7326/m18-0850
22. Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions [Internet]. *arXiv*; 2017 [cited 2025 Jan 10]. Available from: <https://arxiv.org/abs/1705.07874> doi:10.48550/ARXIV.1705.07874
23. Pollock D, Peters MDJ, Khalil H, McInerney P, Alexander L, Tricco AC, et al. Recommendations for the extraction, analysis, and presentation of results in scoping reviews. *JBIM Evid Synth.* 2023;21(3):520–32. doi:10.11124/JBIMES-22-00123

24. Kohl C, McIntosh EJ, Unger S, Haddaway NR, Kecke S, Schiemann J, et al. Online tools supporting the conduct and reporting of systematic reviews and systematic maps: a case study on CADIMA and review of existing tools. *Environ Evid*. 2018;7(1):8. doi:10.1186/s13750-018-0115-5
25. Schemmer M, Kuehl N, Benz C, Bartos A, Satzger G. Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. In: *Proceedings of the 28th International Conference on Intelligent User Interfaces* [Internet]. Sydney NSW Australia: ACM; 2023 [cited 2026 Aug 17]. p. 410–22. Available from: <https://dl.acm.org/doi/10.1145/3581641.3584066> doi:10.1145/3581641.3584066
26. Hasanzadeh F, Josephson CB, Waters G, Adedinsowo D, Azizi Z, White JA. Bias recognition and mitigation strategies in artificial intelligence healthcare applications. *Npj Digit Med*. 2025;8(1):154. doi:10.1038/s41746-025-01503-7
27. Dratsch T, Chen X, Rezazade Mehrizi M, Kloeckner R, Mähringer-Kunz A, Püsken M, et al. Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance. *Radiology*. 2023;307(4):e222176. doi:10.1148/radiol.222176
28. Buçinca Z, Lin P, Gajos KZ, Glassman EL. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In: *Proceedings of the 25th International Conference on Intelligent User Interfaces* [Internet]. Cagliari Italy: ACM; 2020 [cited 2023 Mar 30]. p. 454–64. Available from: <https://dl.acm.org/doi/10.1145/3377325.3377498> doi:10.1145/3377325.3377498
29. Küper A, Lodde GC, Livingstone E, Schadendorf D, Krämer N. Psychological Factors Influencing Appropriate Reliance on AI-enabled Clinical Decision Support Systems: Experimental Web-Based Study Among Dermatologists. *J Med Internet Res*. 2025;27:e58660. doi:10.2196/58660
30. Creswell JW, Plano Clark VL. *Designing and conducting mixed methods research*. Third edition, international student edition. Los Angeles London New Delhi Singapore Washington DC Melbourne: Sage; 2018. 492 p.
31. Stern C, Lizarondo L, Carrier J, Godfrey C, Rieger K, Salmond S, et al. Methodological guidance for the conduct of mixed methods systematic reviews. *JBI Evid Synth*. 2020;18(10):2108–18. doi:10.11124/JBISIR-D-19-00169
32. Sandelowski M, Voils CI, Barroso J. *Defining and Designing Mixed Research Synthesis Studies*. 2006;13(1):29–40.
33. Doshi-Velez F, Kim B. *Towards A Rigorous Science of Interpretable Machine Learning* [Internet]. arXiv; 2017 [cited 2026 Jun 18]. Available from: <http://arxiv.org/abs/1702.08608> doi:10.48550/arXiv.1702.08608
34. Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Wortman Vaughan JW, Wallach H. Manipulating and Measuring Model Interpretability. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* [Internet]. Yokohama Japan: ACM; 2021 [cited 2026 Jun 18]. p. 1–52. Available from: <https://dl.acm.org/doi/10.1145/3411764.3445315> doi:10.1145/3411764.3445315
35. Bansal G, Nushi B, Kamar E, Lasecki WS, Weld DS, Horvitz E. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proc AAAI Conf Hum Comput Crowdsourcing*. 2019;7:2–11. doi:10.1609/hcomp.v7i1.5285
36. Schemmer M, Hemmer P, Kühl N, Benz C, Satzger G. Should I Follow AI-based Advice? Measuring Appropriate Reliance in Human-AI Decision-Making [Internet]. arXiv; 2022 [cited 2024 Apr 11]. Available from: <http://arxiv.org/abs/2204.06916>
37. Bansal G, Wu T, Zhou J, Fok R, Nushi B, Kamar E, et al. 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* [Internet]. Yokohama Japan: ACM; 2021 [cited 2023 Aug 22]. p. 1–16. Available from: <https://dl.acm.org/doi/10.1145/3411764.3445717> doi:10.1145/3411764.3445717
38. Medicines and Healthcare products Regulatory Agency. *Medical devices: post-market surveillance requirements* [Internet]. GOV.UK; 2025. Available from: <https://www.gov.uk/government/publications/medical-devices-post-market-surveillance-requirements>
39. U.S. Food and Drug Administration, Health Canada, Medicines and Healthcare products Regulatory Agency. *Good Machine Learning Practice for Medical Device Development: Guiding Principles*. 2021.
40. U.S. Food and Drug Administration. *Medical Device Reporting (MDR): How to report medical device problems*. 2025. U.S. Food and Drug Administration.
41. *Transparency for Machine Learning-Enabled Medical Devices: Guiding Principles*. U.S. Food and Drug Administration, Health Canada, Medicines and Healthcare products Regulatory Agency.; 2024.
42. Bienefeld N, Boss JM, Lüthy R, Brodbeck D, Azzati J, Blaser M, et al. Solving the explainable AI conundrum by bridging clinicians' needs and developers' goals. *Npj Digit Med*. 2023;6(1):94. doi:10.1038/s41746-023-00837-4
43. Bussone A, Stumpf S, O'Sullivan D. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In: *2015 International Conference on Healthcare Informatics* [Internet]. Dallas, TX, USA: IEEE; 2015 [cited 2024 May 13]. p. 160–9. Available from: <http://ieeexplore.ieee.org/document/7349687/> doi:10.1109/ICHI.2015.26
44. Chanda T, Hauser K, Hobelsberger S, Bucher TC, Garcia CN, Wies C, et al. Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma. *Nat Commun*. 2024;15(1):524. doi:10.1038/s41467-023-43095-4
45. Laxar D, Eitenberger M, Maleczek M, Kaider A, Hammerle FP, Kimberger O. The influence of explainable vs non-explainable clinical decision support systems on rapid triage decisions: a mixed methods study. *BMC Med*. 2023;21(1):359. doi:10.1186/s12916-023-03068-2
46. Sabol P, Sinčák P, Hartono P, Kočan P, Benetinová Z, Blichárová A, et al. Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images. *J Biomed Inform*. 2020;109:103523. doi:10.1016/j.jbi.2020.103523
47. Ackerhans S, Wehkamp K, Petzina R, Dumitrescu D, Schultz C. Perceived Trust and Professional Identity Threat in AI-Based Clinical Decision Support Systems: Scenario-Based Experimental Study on AI Process Design Features. *JMIR Form Res*. 2025;9:e64266. doi:10.2196/64266
48. Rezaeian O, Asan O, Bayrak AE. The impact of AI explanations on clinicians' trust and diagnostic accuracy in breast cancer. *Appl Ergon*. 2025;129:104577. doi:10.1016/j.apergo.2025.104577
49. De Oliveira AC, Azevedo JPC, Ruback L, Moreira R, Teixeira SS, Teles AS. Effect of Explainable Artificial Intelligence on Trust of Mental Health Professionals in an AI-Based System for Suicide Prevention. *IEEE Access*. 2025;13:60987–1005. doi:10.1109/ACCESS.2025.3556245

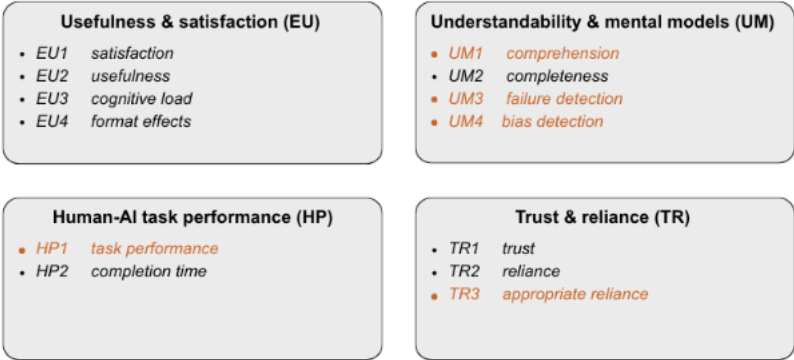
50. Nicolson A, Bradburn E, Gal Y, Papageorgiou AT, Noble JA. The human factor in explainable artificial intelligence: clinician variability in trust, reliance, and performance. *Npj Digit Med.* 2025;8(1):658. doi:10.1038/s41746-025-02023-0
51. Wysocki O, Davies JK, Vigo M, Armstrong AC, Landers D, Lee R, et al. Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. *Artif Intell.* 2023;316:103839. doi:10.1016/j.artint.2022.103839
52. Al-bakri FH, Bejuri WMYW, Al-Andoli MN, Ikram RRR, Khor HM, Sholva Y, et al. A Feature-Augmented Explainable Artificial Intelligence Model for Diagnosing Alzheimer's Disease from Multimodal Clinical and Neuroimaging Data. *Diagnostics.* 2025;15(16):2060. doi:10.3390/diagnostics15162060
53. Antoniadis AM, Galvin M, Heverin M, Wei L, Hardiman O, Mooney C. A Clinical Decision Support System for the Prediction of Quality of Life in ALS. *J Pers Med.* 2022;12(3):435. doi:10.3390/jpm12030435
54. Cabitza F, Campagner A, Natali C, Parimbelli E, Ronzio L, Cameli M. Painting the Black Box White: Experimental Findings from Applying XAI to an ECG Reading Setting. *Mach Learn Knowl Extr.* 2023;5(1):269–86. doi:10.3390/make5010017
55. Naiseh M, Cemiloglu D, Al Thani D, Jiang N, Ali R. Explainable Recommendations and Calibrated Trust: Two Systematic User Errors. *Computer.* 2021;54(10):28–37. doi:10.1109/MC.2021.3076131
56. Famiglioni L, Campagner A, Barandas M, La Maida GA, Gallazzi E, Cabitza F. Evidence-based XAI: An empirical approach to design more effective and explainable decision support systems. *Comput Biol Med.* 2024;170:108042. doi:10.1016/j.combiomed.2024.108042
57. Glick A, Clayton M, Angelov N, Chang J. Impact of explainable artificial intelligence assistance on clinical decision-making of novice dental clinicians. *JAMIA Open.* 2022;5(2):ooac031. doi:10.1093/jamiaopen/ooac031
58. Jabbour S, Fouhey D, Shepard S, Valley TS, Kazerooni EA, Banovic N, et al. Measuring the Impact of AI in the Diagnosis of Hospitalized Patients: A Randomized Clinical Vignette Survey Study. *JAMA.* 2023;330(23):2275. doi:10.1001/jama.2023.22295
59. Das N, Happaerts S, Gyselinck I, Staes M, Derom E, Brussels G, et al. Collaboration between explainable artificial intelligence and pulmonologists improves the accuracy of pulmonary function test interpretation. *Eur Respir J.* 2023;61(5):2201720. doi:10.1183/13993003.01720-2022
60. Gomez C, Smith BL, Zayas A, Unberath M, Canares T. Explainable AI decision support improves accuracy during telehealth strep throat screening. *Commun Med.* 2024;4(1):149. doi:10.1038/s43856-024-00568-x
61. Gombolay GY, Silva A, Schrum M, Gopalan N, Hallman-Cooper J, Dutt M, et al. Effects of explainable artificial intelligence in neurology decision support. *Ann Clin Transl Neurol.* 2024;11(5):1224–35. doi:10.1002/acn3.52036
62. Chen H, Gomez C, Correa ZM, Liu A, Milman T, Eiger-Moscovich M, et al. An interactive and explainable AI approach to improve human-machine teaming in cancer subtyping from digital cytopathology. *Med Image Anal.* 2026;108:103856. doi:10.1016/j.media.2025.103856
63. Gomez C, Wang R, Breininger K, Casey C, Bradley C, Pavlak M, et al. The explainable AI dilemma under knowledge imbalance in specialist AI for glaucoma referrals in primary care. *Npj Digit Med.* 2025;8(1):706. doi:10.1038/s41746-025-02069-0
64. Nagendran M, Festor P, Komorowski M, Gordon AC, Faisal AA. Quantifying the impact of AI recommendations with explanations on prescription decision making. *Npj Digit Med.* 2023;6(1):206. doi:10.1038/s41746-023-00955-z
65. Jacobs M, Pradier MF, McCoy TH, Perlis RH, Doshi-Velez F, Gajos KZ. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Transl Psychiatry.* 2021;11(1):108. doi:10.1038/s41398-021-01224-x
66. Morey DA, Rayo MF. Situated Interpretation and Data: Explainability to Convey Machine Misalignment. *IEEE Trans Hum-Mach Syst.* 2024;54(1):100–9. doi:10.1109/THMS.2023.3334988
67. Sivaraman V, Bukowski LA, Levin J, Kahn JM, Perer A. 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 [Internet]. Hamburg Germany: ACM; 2023 [cited 2025 Sep 11]. p. 1–18. Available from: <https://dl.acm.org/doi/10.1145/3544548.3581075> doi:10.1145/3544548.3581075*
68. Naiseh M, Al-Thani D, Jiang N, Ali R. How the different explanation classes impact trust calibration: The case of clinical decision support systems. *Int J Hum-Comput Stud.* 2023;169:102941. doi:10.1016/j.ijhcs.2022.102941
69. Du Y, Antoniadis AM, McNestry C, McAuliffe FM, Mooney C. The Role of XAI in Advice-Taking from a Clinical Decision Support System: A Comparative User Study of Feature Contribution-Based and Example-Based Explanations. *Appl Sci.* 2022;12(20):10323. doi:10.3390/app122010323
70. Singh A, Jothi Balaji J, Rasheed MA, Jayakumar V, Raman R, Lakshminarayanan V. Evaluation of Explainable Deep Learning Methods for Ophthalmic Diagnosis. *Clin Ophthalmol.* 2021;Volume 15:2573–81. doi:10.2147/OPTH.S312236
71. Bergomi L, Nicora G, Orlowska MA, Podrecca C, Bellazzi R, Fregosi C, et al. Which explanations do clinicians prefer? A comparative evaluation of XAI understandability and actionability in predicting the need for hospitalization. *BMC Med Inform Decis Mak.* 2025;25(1):269. doi:10.1186/s12911-025-03045-0
72. Hur S, Lee Y, Park J, Jeon YJ, Cho JH, Cho D, et al. Comparison of SHAP and clinician friendly explanations reveals effects on clinical decision behaviour. *Npj Digit Med.* 2025;8(1):578. doi:10.1038/s41746-025-01958-8
73. Anjara SG, Janik A, Dunford-Stenger A, Mc Kenzie K, Collazo-Lorduy A, Torrente M, et al. Examining explainable clinical decision support systems with think aloud protocols. Reddy S, editor. *PLOS ONE.* 2023;18(9):e0291443. doi:10.1371/journal.pone.0291443
74. Gaube S, Suresh H, Raue M, Lermer E, Koch TK, Hudecek MFC, et al. Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays. *Sci Rep.* 2023;13(1):1383. doi:10.1038/s41598-023-28633-w
75. E. Ihongbe I, Fouad S, F. Mahmoud T, Rajasekaran A, Bhatia B. Evaluating Explainable Artificial Intelligence (XAI) techniques in chest radiology imaging through a human-centered Lens. Ieracitano C, editor. *PLOS ONE.* 2024;19(10):e0308758. doi:10.1371/journal.pone.0308758
76. Barda AJ, Horvat CM, Hochheiser H. A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. *BMC Med Inform Decis Mak.* 2020;20(1):257. doi:10.1186/s12911-020-01276-x

77. Corti L, Oltmans R, Jung J, Balayn A, Wijsenbeek M, Yang J. It Is a Moving Process": Understanding the Evolution of Explainability Needs of Clinicians in Pulmonary Medicine. In: Proceedings of the CHI Conference on Human Factors in Computing Systems [Internet]. Honolulu HI USA: ACM; 2024 [cited 2024 Sep 23]. p. 1–21. Available from: <https://dl.acm.org/doi/10.1145/3613904.3642551> doi:10.1145/3613904.3642551
78. Guo G, Deng L, Tandon A, Endert A, Kwon BC. MiMICRI: Towards Domain-centered Counterfactual Explanations of Cardiovascular Image Classification Models. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency [Internet]. Rio de Janeiro Brazil: ACM; 2024 [cited 2024 Sep 23]. p. 1861–74. Available from: <https://dl.acm.org/doi/10.1145/3630106.3659011> doi:10.1145/3630106.3659011
79. Gutiérrez F, Htun NN, Vanden Abeele V, De Croon R, Verbert K. Explaining Call Recommendations in Nursing Homes: a User-Centered Design Approach for Interacting with Knowledge-Based Health Decision Support Systems. In: 27th International Conference on Intelligent User Interfaces [Internet]. Helsinki Finland: ACM; 2022 [cited 2024 Sep 23]. p. 162–72. Available from: <https://dl.acm.org/doi/10.1145/3490099.3511158> doi:10.1145/3490099.3511158
80. Kovalchuk SV, Kopanitsa GD, Derevitskii IV, Matveev GA, Savitskaya DA. Three-stage intelligent support of clinical decision making for higher trust, validity, and explainability. *J Biomed Inform.* 2022;127:104013. doi:10.1016/j.jbi.2022.104013
81. Schwartz JM, George M, Rossetti SC, Dykes PC, Minshall SR, Lucas E, et al. Factors Influencing Clinician Trust in Predictive Clinical Decision Support Systems for In-Hospital Deterioration: Qualitative Descriptive Study. *JMIR Hum Factors.* 2022;9(2):e33960. doi:10.2196/33960
82. Panigutti C, Beretta A, Fadda D, Giannotti F, Pedreschi D, Perotti A, et al. Co-design of Human-centered, Explainable AI for Clinical Decision Support. *ACM Trans Interact Intell Syst.* 2023;13(4):1–35. doi:10.1145/3587271
83. Mandava R, Dalavai L, Vellela SS, Purimetla NR, Mohan BK, Harinadh T. An In-Depth Study on the Integration of Explainable AI Techniques to Enhance Interpretability in Clinical Risk Prediction Models. In: 2025 Second International Conference on Networks and Soft Computing (ICNSoC) [Internet]. Vadlamudi, India: IEEE; 2025 [cited 2026 Jun 5]. p. 43–7. Available from: <https://ieeexplore.ieee.org/document/11344116/> doi:10.1109/ICNSoC66817.2025.00021
84. Harari R, Al-Taweel A, Ahram T, Shokoohi H. Explainable AI and Augmented Reality in Transesophageal Echocardiography (TEE) Imaging. In: 2024 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR) [Internet]. Los Angeles, CA, USA: IEEE; 2024 [cited 2024 Sep 23]. p. 306–9. Available from: <https://ieeexplore.ieee.org/document/10445572/> doi:10.1109/AIxVR59861.2024.00051
85. Nagendran M, Festor P, Komorowski M, Gordon AC, Faisal AA. Eye tracking insights into physician behaviour with safe and unsafe explainable AI recommendations. *Npj Digit Med.* 2024;7(1):202. doi:10.1038/s41746-024-01200-x
86. Kyrimi E, Mossadegh S, Tai N, Marsh W. An incremental explanation of inference in Bayesian networks for increasing model trustworthiness and supporting clinical decision making. *Artif Intell Med.* 2020;103:101812. doi:10.1016/j.artmed.2020.101812
87. Martinez-Villasenor L, Ponce H, Martinez-Velasco A, Miralles-Pechuan L. An Explainable Tool to Support Age-related Macular Degeneration Diagnosis. In: 2022 International Joint Conference on Neural Networks (IJCNN) [Internet]. Padua, Italy: IEEE; 2022 [cited 2024 Sep 23]. p. 1–8. Available from: <https://ieeexplore.ieee.org/document/9892895/> doi:10.1109/IJCNN55064.2022.9892895
88. Zou L, Goh HL, Liew CJY, Quah JL, Gu GT, Chew JJ, et al. Ensemble Image Explainable AI (XAI) Algorithm for Severe Community-Acquired Pneumonia and COVID-19 Respiratory Infections. *IEEE Trans Artif Intell.* 2023;4(2):242–54. doi:10.1109/TAI.2022.3153754
89. Barr Kumarakulasinghe N, Blomberg T, Liu J, Saraiva Leao A, Papapetrou P. Evaluating Local Interpretable Model-Agnostic Explanations on Clinical Machine Learning Classification Models. In: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS) [Internet]. Rochester, MN, USA: IEEE; 2020 [cited 2024 Sep 23]. p. 7–12. Available from: <https://ieeexplore.ieee.org/document/9183167/> doi:10.1109/CBMS49503.2020.00009
90. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. *J Am Med Inform Assoc.* 2017;24(2):423–31. doi:10.1093/jamia/ocw105
91. Lipshitz R, Klein G, Orasanu J, Salas E. Taking stock of naturalistic decision making. *J Behav Decis Mak.* 2001;14(5):331–52. doi:10.1002/bdm.381
92. Abbas Q, Jeong W, Lee SW. Explainable AI in Clinical Decision Support Systems: A Meta-Analysis of Methods, Applications, and Usability Challenges. *Healthcare.* 2025;13(17):2154. doi:10.3390/healthcare13172154
93. Agha-Mir-Salim L, Mosch L, Klopfenstein SAI, Wunderlich MM, Frey N, Poncette AS, et al. Artificial Intelligence Competencies in Postgraduate Medical Training in Germany. In: Séroussi B, Weber P, Dhombres F, Grouin C, Liebe JD, Pelayo S, et al., editors. *Studies in Health Technology and Informatics* [Internet]. IOS Press; 2022 [cited 2025 Feb 19]. Available from: <https://ebooks.iospress.nl/doi/10.3233/SHTI220589> doi:10.3233/SHTI220589
94. Babic B, Gerke S, Evgeniou T, Cohen IG. Beware explanations from AI in health care. *Science.* 2021;373(6552):284–6. doi:10.1126/science.abg1834
95. Marques-Silva J, Huang X. Explainability Is Not a Game. *Commun ACM.* 2024;67(7):66–75. doi:10.1145/3635301
96. Huang X, Marques-Silva J. On the failings of Shapley values for explainability. *Int J Approx Reason.* 2024;171:109112. doi:10.1016/j.ijar.2023.109112
97. Kumar IE, Venkatasubramanian S, Scheidegger C, Friedler SA. Problems with Shapley-value-based explanations as feature importance measures. 2020.
98. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity Checks for Saliency Maps. 2018.
99. Schrills T, Franke T, Hoesterey S, Roesler E. Questioning trust in AI research: exploring the influence of trust assessment on dependence in AI-assisted decision-making. *Behav Inf Technol.* 2026;45(8):1644–60. doi:10.1080/0144929X.2025.2553153
100. Brenner PS, DeLamater J. Lies, Damned Lies, and Survey Self-Reports? Identity as a Cause of Measurement Bias. *Soc Psychol Q.* 2016;79(4):333–54. doi:10.1177/0190272516628298
101. Schemmer M, Kühl N, Benz C, Bartos A, Satzger G. Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations [Internet]. 2023 [cited 2026 Jun 10]. Available from: <http://arxiv.org/abs/2302.02187> doi:10.1145/3581641.3584066

102. Chromik M, Eiband M, Buchner F, Krüger A, Butz A. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In: 26th International Conference on Intelligent User Interfaces [Internet]. College Station TX USA: ACM; 2021 [cited 2026 Jun 10]. p. 307–17. Available from: <https://dl.acm.org/doi/10.1145/3397481.3450644> doi:10.1145/3397481.3450644
103. Frey N, Agha-Mir-Salim L, Hinz E, Poncette AS, Balzer F. Assessing Healthcare Stakeholder Understanding of Machine Learning Documentation. In: Andrikopoulou E, Gallos P, Arvanitis TN, Austin R, Benis A, Cornet R, et al., editors. Studies in Health Technology and Informatics [Internet]. IOS Press; 2025 [cited 2026 May 18]. Available from: <https://ebooks.iospress.nl/doi/10.3233/SHTI250336> doi:10.3233/SHTI250336
104. Agha-Mir-Salim L, Fehlhaber AL, Frey N, Krumpal I, Näher AF. 1 Addressing Algorithmic Bias Through Data Disclosure: 2 Evidence from a Randomized Experiment Among Physicians.
105. Vasconcelos H, Jörke M, Grunde-McLaughlin M, Gerstenberg T, Bernstein MS, Krishna R. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc ACM Hum-Comput Interact.* 2023;7(CSCW1):1–38. doi:10.1145/3579605
106. Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In: *Advances in Psychology* [Internet]. Elsevier; 1988 [cited 2024 Jan 27]. p. 139–83. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0166411508623869> doi:10.1016/S0166-4115(08)62386-9
107. Ehsan U, Riedl MO. Explainability pitfalls: Beyond dark patterns in explainable AI. *Patterns.* 2024;5(6):100971. doi:10.1016/j.patter.2024.100971
108. Hatherley J, Sparrow R, Howard M. The Virtues of Interpretable Medical Artificial Intelligence. *Camb Q Healthc Ethics.* 2022;1–10. doi:10.1017/S0963180122000305
109. Arun N, Gaw N, Singh P, Chang K, Aggarwal M, Chen B, et al. Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging. *Radiol Artif Intell.* 2021;3(6):e200267. doi:10.1148/ryai.2021200267
110. DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat Mach Intell.* 2021;3(7):610–9. doi:10.1038/s42256-021-00338-7
111. Geirhos R, Jacobsen JH, Michaelis C, Zemel R, Brendel W, Bethge M, et al. Shortcut learning in deep neural networks. *Nat Mach Intell.* 2020;2(11):665–73. doi:10.1038/s42256-020-00257-z
112. Wilming R, Kieslich L, Clark B, Haufe S. Theoretical Behavior of XAI Methods in the Presence of Suppressor Variables.
113. Lyu Q, Apidianaki M, Callison-Burch C. Towards Faithful Model Explanation in NLP: A Survey [Internet]. *arXiv*; 2024 [cited 2026 Aug 27]. Available from: <http://arxiv.org/abs/2209.11326> doi:10.48550/arXiv.2209.11326
114. Clark B, Oliveira M, Wilming R, Haufe S. Feature salience - not task-informativeness - drives machine learning model explanations [Internet]. *arXiv*; 2026 [cited 2026 Aug 27]. Available from: <http://arxiv.org/abs/2602.09238> doi:10.48550/arXiv.2602.09238
115. Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, et al. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV) [Internet]. *arXiv*; 2018 [cited 2026 Jun 19]. Available from: <http://arxiv.org/abs/1711.11279> doi:10.48550/arXiv.1711.11279
116. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med.* 2022;28(5):924–33. doi:10.1038/s41591-022-01772-9
117. Guo Z, Wu Y, Hartline JD, Hullman J. A Decision Theoretic Framework for Measuring AI Reliance. In: *The 2024 ACM Conference on Fairness, Accountability, and Transparency* [Internet]. Rio de Janeiro Brazil: ACM; 2024 [cited 2026 Jul 22]. p. 221–36. Available from: <https://dl.acm.org/doi/10.1145/3630106.3658901> doi:10.1145/3630106.3658901
118. Agha-Mir-Salim L, Frey N, Kaiser Z, Mosch L, Weicken E, Freyer O, et al. The Clinician Model Card: development and evaluation of clinician-centered documentation for AI-based clinical decision support [Internet]. *Health Informatics*; 2026 [cited 2026 May 22]. Available from: <http://medrxiv.org/lookup/doi/10.64898/2026.04.15.26350930> doi:10.64898/2026.04.15.26350930
119. Boie SD, Reis F, Frey N, Grünewald E, Balzer F. Calibration of Self-Reported Confidence and Accuracy of Large Language Models in Medical Question Answering. *J Med Syst.* 2026;50(1):103. doi:10.1007/s10916-026-02430-0
120. Frey N, Giesa N, Agha-Mir-Salim L, Balzer F. Physicians’ Perspectives on Predictive Uncertainty in Machine Learning Models. In: Giacomini M, Delgado J, Arvanitis TN, Andrikopoulou E, Benis A, Balestra G, et al., editors. Studies in Health Technology and Informatics [Internet]. IOS Press; 2026 [cited 2026 Jun 2]. Available from: <https://ebooks.iospress.nl/doi/10.3233/SHTI260277> doi:10.3233/SHTI260277

Figures

CONSTRUCT AXIS - What was evaluated?



OPERATIONALIZATION-CATEGORY AXIS - How was it operationalized?

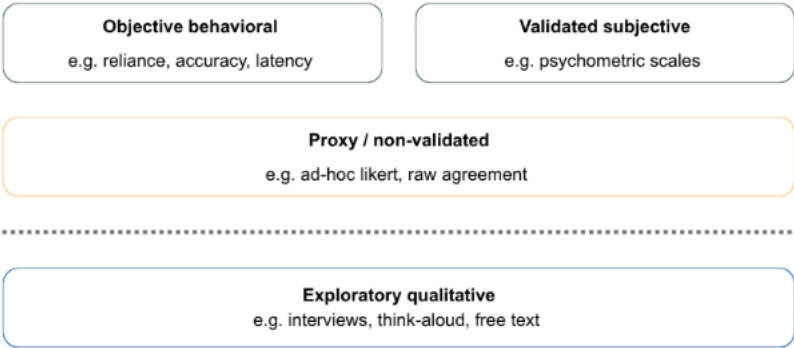


Figure 1

Human-centered XAI evaluation framework adapted from Mohseni et al.[15]. The construct axis comprises four dimensions—explanation usefulness and satisfaction, understandability and mental models, trust and reliance, and human–AI task performance—and their subdimensions. Orange marks constructs containing a safety-relevant oversight capacity; for UM1, this refers specifically to criterion-scored objective model-behavior comprehension rather than to all understandability mappings. The operationalization-category axis comprises objective behavioral, validated subjective, proxy/non-validated, and exploratory qualitative evidence.

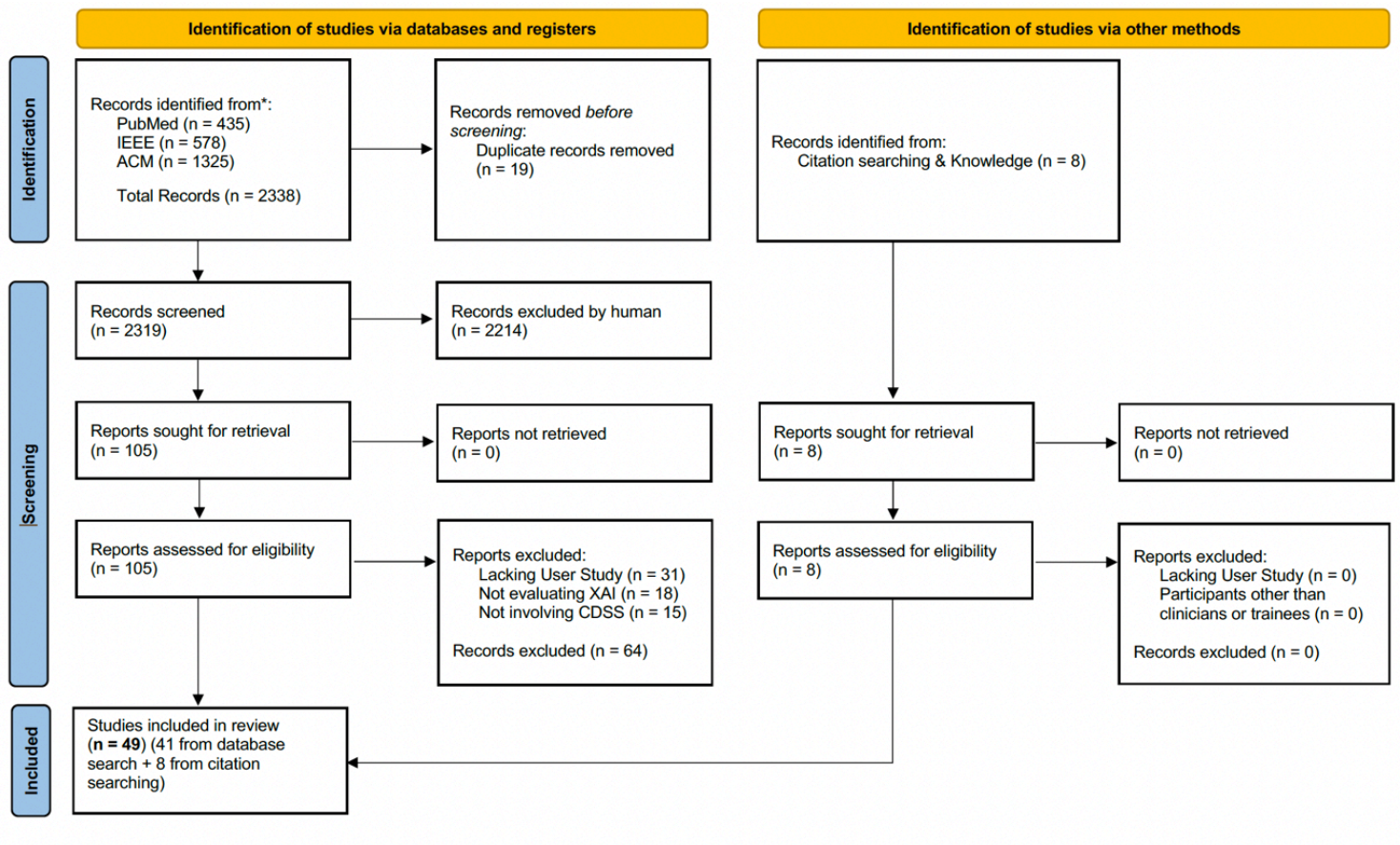


Figure 2

Study selection process.

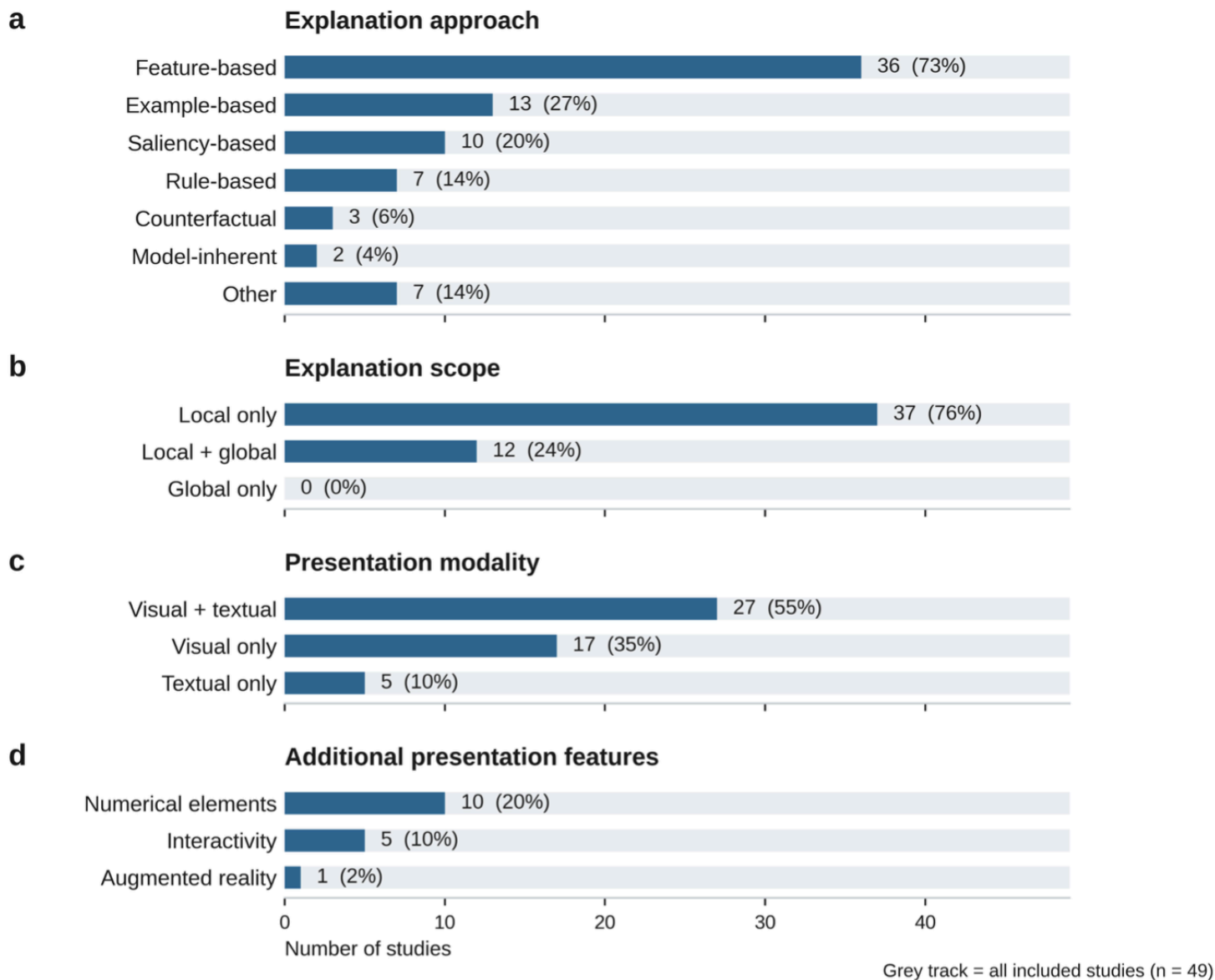


Figure 3

Explanation characteristics of the 49 included studies. Bars show the number of studies; grey tracks indicate the full corpus (n = 49). (a) Explanation approach; multi-label, as a single study could combine several approaches (sum = 78). "Model-inherent" denotes intrinsically interpretable models rather than post-hoc explanation methods; "Other" comprises contextual or workflow-based, narrative, confidence-based, contrastive or outcome-based, heuristic, and possibility-based explanations (one study each). (b) Explanation scope; mutually exclusive. (c) Presentation modality; mutually exclusive. (d) Additional presentation features; multi-label and independent of modality: numerical elements comprise probabilities, confidence values, or risk decompositions; interactivity denotes user-manipulable displays; augmented reality denotes overlay-based presentation. Named XAI techniques and their frequencies are reported in Supplementary Results Table A2.4.

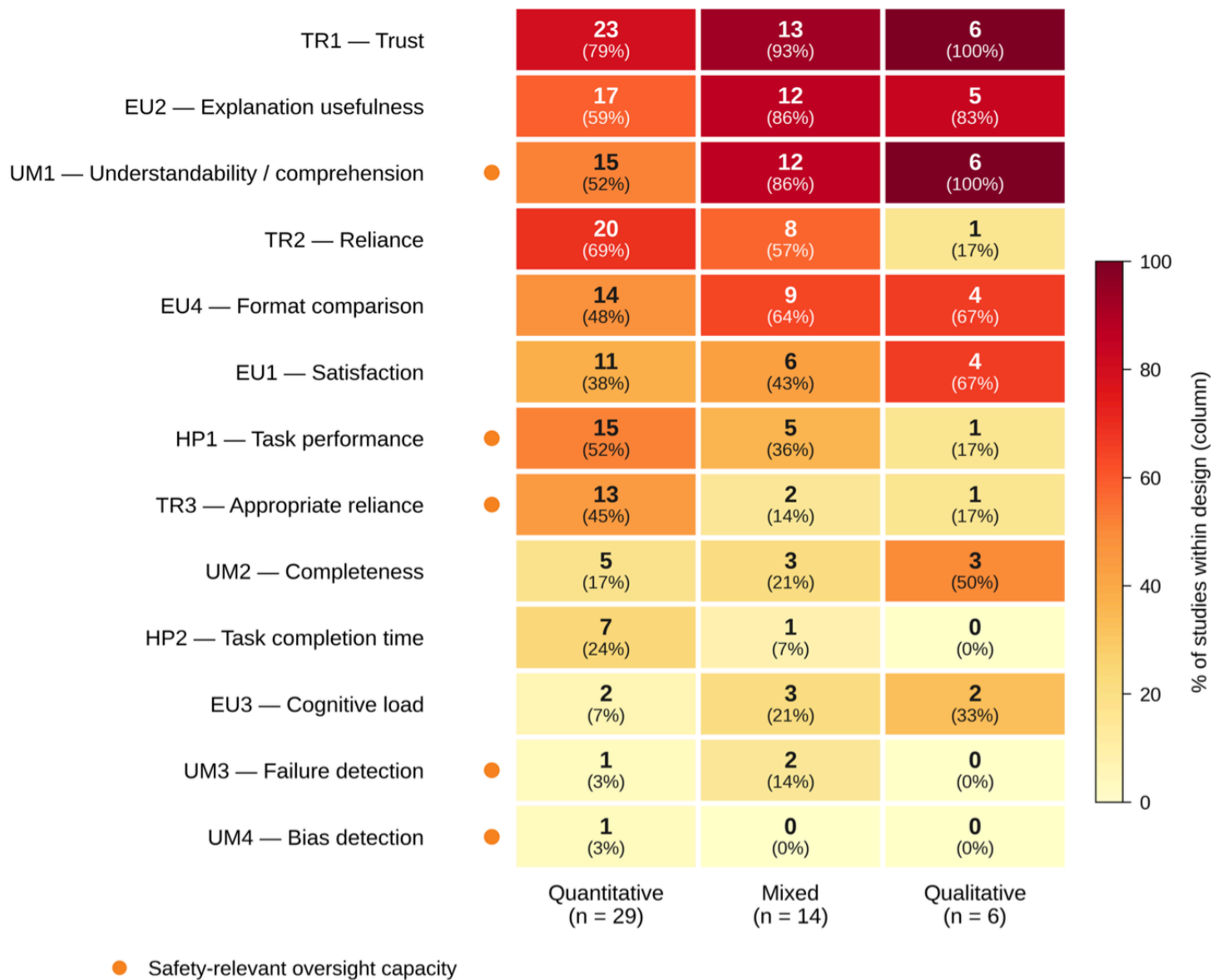


Figure 4

Construct coverage by study design. Cells show the number of studies assessing each construct within quantitative, mixed-methods ("Mixed"), and qualitative designs; shading indicates the percentage of studies within each design column. Counts are multi-label because individual studies could assess several constructs. Orange dots mark the five safety-relevant oversight capacities defined in the Methods; for UM1, the marker refers specifically to criterion-scored objective model-behavior comprehension.

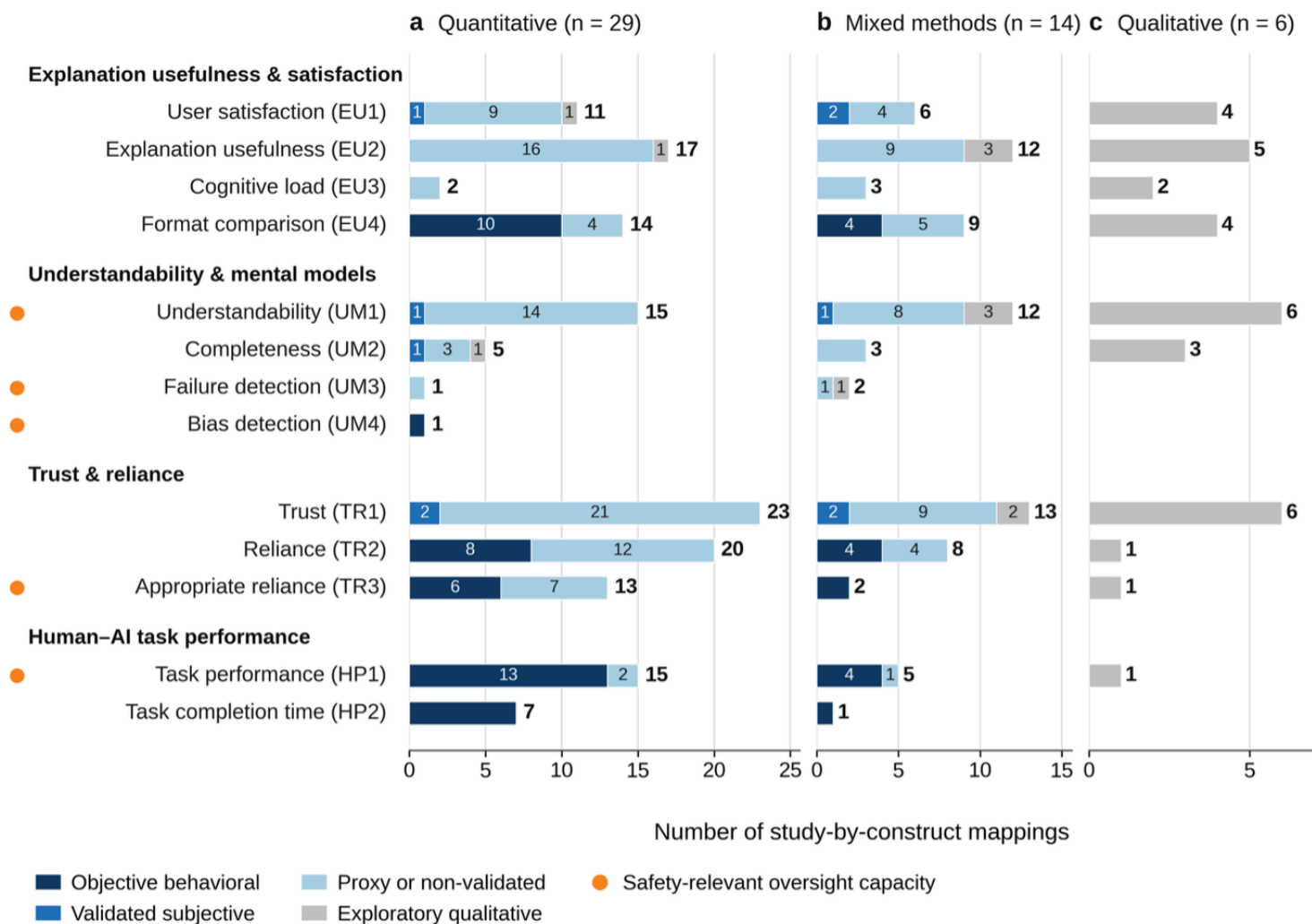


Figure 5

Operationalization categories by construct, stratified by study design. Each bar shows the number of study-by-construct mappings for one construct, decomposed by operationalization category: objective behavioral, validated subjective, proxy or non-validated, and exploratory qualitative. (a) Quantitative studies (n = 29; 144 mappings), (b) mixed-methods studies (n = 14; 76 mappings), and (c) qualitative studies (n = 6; 33 mappings), where all mappings were classified as exploratory qualitative evidence. Counts are multi-label because individual studies could assess several constructs. Orange dots indicate the five safety-relevant oversight capacities defined in the Methods; for UM1, the marker refers specifically to criterion-scored objective model-behavior comprehension rather than to all UM1 mappings.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [RADAR7XAIscopingSupplementary.docx](#)
- [RADAR7XAIscopingSupplementaryResults.docx](#)