Evaluating Transparency in the Development of Artificial Intelligence Systems: A Systematic Literature Review

Giulia Karanxha, Paulinus Ofem
Business and Information Technology, Laurea University of Applied Sciences, Espoo, Finland

***Abstract*—Transparency is increasingly recognised as a cornerstone of trustworthy artificial intelligence (AI), yet its operationalisation remains fragmented and underdeveloped. Existing methods often rely on qualitative checklists or domain-specific case studies, limiting comparability, reproducibility, and regulatory alignment. This paper presents a Systematic Literature Review (SLR) of 28 peer-reviewed studies that explicitly propose or apply methods for evaluating transparency in AI systems (2019-July 2025). The review identifies recurring themes such as traceability, explainability, and communication, and classifies evaluation approaches by metric type and calculation type. Empirically, checklist-based instruments are the most frequent evaluation form (9/28, 32%), followed by scenario-based qualitative assessments (5/28, 18%). Most (9/28, 32%) research on AI applications occurs in healthcare; references to legal or ethical frameworks appear in 19/28 studies (67%), although traceable mappings to specific obligations are rare. The results of the quality assessment highlight strengths in methodological clarity, but reveal persistent gaps in benchmarking, stakeholder inclusion, and lifecycle integration. Based on these findings, this study informs the adaptation of the Z-Inspection® process within the context of AI development projects and motivates a Transparency Artefact Registry (TAR), a structured, metadata-based mechanism for capturing and reusing transparency artefacts across system lifecycles. By embedding transparency evaluation into AI development workflows, the proposed approach seeks to provide verifiable, repeatable, and regulation-aligned practices for assessing transparency in complex AI systems.**

Keywords*—Artificial intelligence; transparency evaluation; trustworthy AI; transparency metrics; EU AI Act; systematic literature review

I. Introduction

The rapid deployment of artificial intelligence (AI) in high-impact domains has intensified calls for transparency as a cornerstone of Trustworthy AI. Transparency enables stakeholders to understand, contextualise, and, where necessary, contest AI-driven decisions [1]. Its importance is underscored by the EU Artificial Intelligence Act (Regulation (EU) 2024/1689), which mandates that high-risk AI systems and general-purpose AI models provide features such as explainability, traceability, and clear user information [2]. Yet, despite these obligations, the operationalisation of transparency remains fragmented, with varying definitions, scopes, and evaluation practices across disciplines and application domains. This poor clarity risks undermining both accountability and compliance in real-world AI deployments.

For the MANOLO Project (Trustworthy Efficient AI for Cloud-Edge Computing) [3], this challenge presents an opportunity to explore further. This project aims to develop a complete set of trustworthy algorithms and tools that would enable AI systems to achieve better efficiency and optimise their operations on the edge of the cloud continuum and in sectors such as healthcare, telecommunications, and robotics, where systems must combine performance and energy requirements with strict ethical and legal compliance. In such contexts, the absence of structured, measurable approaches to transparency achievement and evaluation poses not only a compliance risk but also a practical barrier to demonstrating accountability to regulators, end-users, and other stakeholders.

To address this, we conducted a Systematic Literature Review (SLR) of peer-reviewed studies published between January 2019 and July 2025. The review identifies how transparency in AI is defined, measured, and operationalised, and where gaps remain. By mapping existing approaches and their limitations, the SLR provides the empirical and conceptual foundation needed for AI projects to adopt a structured transparency evaluation strategy.

This review also serves as the basis for further methodological contributions. Its findings directly inform our conference paper, which introduced an artefact-driven approach to transparency evaluation in applied AI contexts. In this vein, this paper is not an isolated study but the first step in a broader research programme: establishing a systematic evidence base for transparency evaluation, and subsequently proposing solutions that can be embedded into frameworks such as Z-Inspection® [4] to support continuous, auditable, and stakeholder-relevant transparency in AI projects and beyond.

The remainder of this paper is organised as follows. Motivation situates the work in the MANOLO and Z-Inspection® contexts; Related Work maps policy, governance, and SE/IS perspectives; Methodology details the SLR protocol (scope, PICOC, search, screening, quality appraisal, extraction); Results report the evidence base and descriptive statistics; Thematic Analysis synthesises cross-cutting patterns; Discussion provides implications of findings and advantages over prevalent approaches; Conclusion summarises limitations and future work.

Contributions and roadmap. This study 1) focuses on operational, metric-based transparency *evaluation* methods rather than principles, 2) proposes a compact comparative taxonomy (metric types) and quantifies patterns across 28 studies, 3) offers a cross-domain synthesis bridging AI governance and SE/IS consideration of transparency as a first class non-functional requirement, and 4) links identified gaps to an

artefact-driven solution space (Transparency Artefact Registry, TAR) aligned with Z-Inspection® and the EU Artificial Intelligence Act (EU AI Act). Section II motivates the challenge of evaluating transparency in AI development projects such as the Manolo project; Section III synthesises related work; Section IV details the SLR protocol; Section V reports descriptive results; Section IV-G provides thematic analysis; Section VII discusses implications; Section VIII concludes.

II. Motivation

Within MANOLO, the evaluation of AI trustworthiness is grounded in the Z-Inspection® framework [4], a holistic process for assessing AI's ethical compliance, technical robustness, and legal alignment. Although Z-Inspection® provides comprehensive coverage of the trustworthiness dimensions, it does not offer a structured, dedicated methodology for the evaluation of transparency throughout the AI lifecycle.

For AI projects that utilise Z-Inspection®, this limitation creates both a compliance risk under the EU AI Act and a practical challenge to demonstrate transparency and accountability to stakeholders. Therefore, a systematic investigation of existing transparency evaluation methods was required to establish a structured evidence base. This SLR responds to this need by synthesising how transparency has been defined, measured, and operationalised in peer-reviewed studies. In this way, the review provides the foundation for future AI projects that will adapt Z-Inspection® with measurable transparency criteria aligned with regulatory obligations and stakeholder needs.

III. Related Work

Transparency is increasingly recognised as a key requirement for ethical and trustworthy AI, yet its practical achievement and evaluation remain challenging. Over 94% of ethical AI guidelines include transparency or explainability as core principles [5], but few provide actionable tools for assessment. This gap between normative commitments and operational practices has driven attention from regulators, practitioners, and researchers alike.

A. Policy and Regulatory Frameworks

The concept of transparency gained formal traction in the European policy landscape through the High-Level Expert Group on AI (AI HLEG), which identified transparency, encompassing traceability, explainability, and user communication, as one of seven requirements for Trustworthy AI [1]. While influential in shaping the EU AI Act [2], these guidelines are not legally binding. The Act introduces layered obligations for high-risk AI systems (Article 13), general-purpose AI models, and AI systems that interact with humans (Article 50), but leaves their operationalisation open to interpretation [6].

Sector-specific initiatives further illustrate this shift from high-level principles to more practical instruments. The TITAN Guideline [7] offers a checklist for AI disclosure in scholarly publishing, while Genovesi et al. [8] compare how transparency is addressed in standards and policy artefacts. Although valuable, such approaches remain largely domain-bound and rarely demonstrate applicability across diverse AI systems.

B. Societal Accountability and Governance

Beyond regulatory discourse, concerns persist about the societal implications of opaque AI. Busuioc [9] underscores the risks of algorithmic systems becoming embedded in the public sector while remaining resistant to challenge, thereby undermining institutional accountability. This perspective positions transparency not only as a compliance issue but also as a prerequisite for democratic oversight. In line with this, Panigutti et al. [10] argue that transparency must be approached as a dynamic, stakeholder-aware practice rather than a static disclosure requirement. These insights highlight the broader governance role of transparency, which extends beyond technical explainability.

C. Software Engineering and Information Systems Foundations

Long before the rise of AI governance, transparency was conceptualised within software engineering (SE) and information systems (IS) as a non-functional requirement (NFR). Leite and Cappelli [11] introduced transparency as a "soft goal," supported by attributes such as accessibility, auditability, and usability. Hosseini et al. [12] extended this by framing transparency as a bidirectional process between information providers and receivers, thereby emphasising the importance of communication and user understanding. Spagnuelo et al. [13] operationalised the concept further by proposing sector-specific metrics for auditability, accountability, and verifiability in IT systems.

Building on these conceptual foundations, Ofem and colleagues advanced transparency in software engineering through a sequence of works. Their 2022 systematic review [14] clarified the fragmented definitions of transparency and positioned it explicitly as a non-functional requirement, comparable to established quality attributes such as reliability or usability. In the same year, Isong, Ofem, and Lugayizi [15] proposed a framework that operationalises transparency via quality factors, such as accessibility, usability, understandability, modifiability, and reusability, each linked to artefact- and process-level indicators. While these factors provide a structured starting point, the framework acknowledges that specific instantiations may be domain-dependent and must be adapted to stakeholder requirements. This line of work culminated in the 2024 metrics study [16], which introduced empirically validated measurement instruments and an improvement model, thereby transforming transparency from an abstract principle into a set of actionable, lifecycle-aware practices. Although developed in the SE domain, these contributions are directly relevant to AI governance, where similar challenges of definition, measurement, and lifecycle integration remain unresolved.

D. Relevance to AI Transparency Evaluation

Although originating in SE and IS, these approaches are directly relevant to AI governance. The challenges of lifecycle integration, stakeholder involvement, and continuous evaluation are not unique to AI but are shared with complex software systems more broadly. Yet, AI-focused frameworks have thus far incorporated these traditions only partially. Most emphasise explainability or documentation, but rarely adopt SE/IS practices such as artefact registries, maturity models, or quality factor-metric mappings.

This review therefore situates AI transparency evaluation within a broader trajectory: from ethical and regulatory guidelines, through societal accountability concerns, to structured SE/IS methodologies. While progress has been made in translating abstract principles into operational criteria, most existing efforts remain either domain-specific, non-standardised, or lacking in mechanisms for continuous lifecycle monitoring. The present SLR contributes to addressing that gap by systematically identifying and analysing peer-reviewed studies that explicitly propose or apply methods for evaluating transparency in AI systems, and by assessing their relationship to established SE/IS traditions.

1) Integrative synthesis: Policy and governance sources prioritise obligations, rights, and auditability, which have produced tools that mostly take the form of checklists and high-level controls. In contrast, SE/IS literature frames transparency as a non-functional requirement and develops artefact and process-level indicators with measurable quality factors. These viewpoints converge around traceability and documentation, taking into account logs, records, and provenance as core evidence, while also acknowledging lifecycle concerns. These perspectives diverge in emphasis. AI governance often treats explainability as a proxy for transparency, while SE/IS considers transparency as a concept with measurable quality constructs. Methodological overlaps lack coherence: many governance-oriented frameworks document evidence without scoring, whereas several SE/IS approaches score internal artefacts yet rarely assess whether information is communicated to stakeholders in clear, timely, and usable ways (e.g., plain-language notices, model cards, user-tested interfaces). Taken together, this reveals the gap this SLR targets: the field lacks a standardised, lifecycle-grounded, cross-domain evaluation scheme that connects regulatory obligations to measurable artefacts and metrics.

IV. Methodology

This study adopts a SLR approach to investigate how transparency is evaluated in artificial intelligence (AI) systems. The review is guided by the following research questions:

- RQ1: What are the existing approaches for implementing and evaluating transparency in AI systems?
- RQ2: How do these approaches address key transparency concerns?
- RQ3: What are the strengths, limitations, and gaps of these approaches?
- RQ4: How applicable are these approaches in real-world AI systems?

The aim is to identify methodological trends, assess sectoral and regulatory relevance, and support the development of actionable stakeholder-centred transparency evaluation strategies.

TABLE I. Total Search Results Summary

Engine	Total	Reviewed	Overlap
Google Scholar	45,700	179	9
IEEE Xplore	1,799	83	–
Web of Science	4,341	104	22
ScienceDirect	3,998	105	2
ACM DL	4,078	71	7

A. Scope Framing and Objectives

Transparency was examined as a multidimensional concept that encompasses traceability, communication, and explainability [17]. The purpose of the review was to synthesise existing evaluation methods, identify common practices and limitations, and provide a structured basis for future development of transparency metrics and frameworks. The review focused on peer-reviewed literature published between January 2019 and July 2025 to reflect the growing influence of transparency in AI governance discourse, especially in light of regulatory developments such as the EU AI Act (2024) [2].

B. PICOC Structure

The scope of the review was defined using the PICOC model.

- Population: AI systems deployed across various sectors and domains.
- Intervention: Methods, frameworks, or metrics introduced or applied to evaluate transparency.
- Comparison: different approaches for implementing and evaluating transparency.
- Outcome: Assessment of the strengths, limitations, and applicability of approaches to AI development projects.
- Context: AI development projects.

This framing supported a focused yet cross-sectoral exploration of how transparency is operationalised and assessed in contemporary AI research.

C. Search Strategy

The literature search was conducted between January 2019 and July 2025 across five major academic databases: Google Scholar, IEEE Xplore, Web of Science, ScienceDirect, and the ACM Digital Library. These sources were selected for their broad coverage of AI, computer science, ethics, and governance literature. The consolidated counts are summarised in Table I. Search queries combined keywords related to transparency, artificial intelligence, and evaluation, adapted to the syntax of each platform. An example query was:

```
("Artificial intelligence system" OR "AI
        software") AND
```

```
("AI transparency" OR "transparent AI" OR
"trustworthy AI") AND
("transparency evaluation" OR
"transparency metrics" OR "transparency
framework" OR "trustworthy framework" OR
"assessment")
```

Boolean operators and filters (e.g., English language, peer-reviewed) were applied where supported. A heuristic stopping rule was used to manage feasibility, with searches concluded when ten consecutive irrelevant results were encountered. A total of over 59,000 results were screened, including 45,700 from Google Scholar, 4,341 from Web of Science, 1,799 from IEEE, 4,078 from ACM, and 3,998 from ScienceDirect.

D. Screening and Eligibility (PRISMA)

Screening followed the PRISMA 2020 framework. The selection and exclusion process is summarised in the PRISMA flow diagram shown in Fig. 1. After duplicate removal and title/abstract screening, 542 articles were assessed. A total of 28 peer-reviewed studies were included in the final review. The inclusion and exclusion criteria were:

- Inclusion:
  - Peer-reviewed journal or conference publications
  - Published between January 2019 and July 2025
  - Explicitly introduced or applied a method, framework, or metric to evaluate transparency in AI systems
- Exclusion:
  - Discussed transparency only conceptually, without operational methods or proposed governance mechanisms
  - Focused solely on explainability, fairness, or accountability without reference to transparency
  - Non-English or inaccessible full texts

E. Quality Assessment

The included studies were evaluated using a five-criterion quality checklist, adapted from a prior systematic literature review in software engineering [14]:

- Is the evaluation method explicitly described?
- Are structured criteria or metrics provided?
- Are the strengths and limitations discussed?
- Is the method compared with other frameworks?
- Is the approach applied in real-world AI use cases?

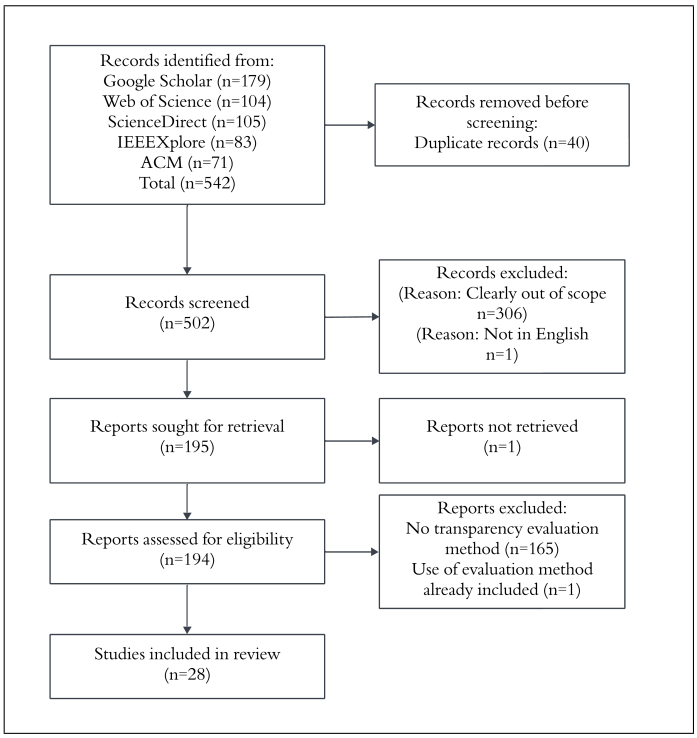

Fig. 1. PRISMA 2020 screening flow for the review window January 2019-July 2025.

Each item was scored as Yes (1), Partly (0.5), or No (0). Studies were not excluded based on their quality score, but scores were used descriptively to support analysis. As shown in Table II, most studies scored 4 or above out of 5.

F. Data Extraction and Classification

A structured extraction template was applied to each study, capturing:

- Transparency definition and theoretical framing.
- Name and structure of evaluation framework or method.
- Metric type (e.g., checklist, score, index, rating, qualitative, survey, explainability-based).
- Metric calculation type (e.g., single score, ordinal classification, aggregated, scenario-based).
- Sector or domain of application.
- Alignment with legal or ethical frameworks (e.g., GDPR, EU AI Act).
- Reported strengths and limitations.

Metrics were classified using a two-dimensional taxonomy: metric type and calculation type. Where multiple techniques

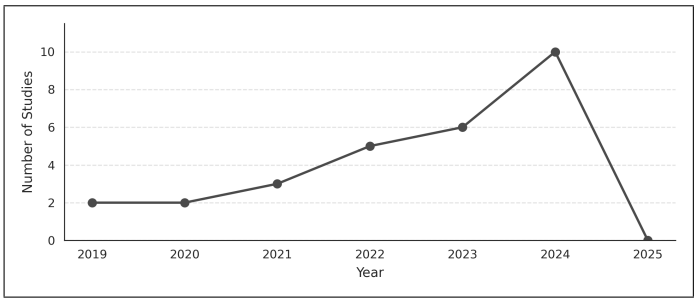

Fig. 2. Included studies per year (2019-2025; cut-off July 2025), n=28.

were reported, the dominant method was selected based on its evaluative function.

1) Data management: Extracted variables were recorded in a Microsoft Excel workbook, which constitutes the coded analytic dataset used for the descriptive analyses and quality appraisal.

2) Dataset: No experimental or task datasets were analysed. The only data used in this study are the 28 included papers (2019-July 2025) and the coded Excel extraction sheet derived from them. The Excel sheet will be made available upon request.

G. Thematic Analysis

Thematic synthesis followed Braun and Clarke's method [18]. Codes were generated from extracted data fields and grouped inductively into higher-level themes. These themes reflect patterns in how transparency is conceptualised and evaluated across the selected studies. The analysis served to uncover recurrent gaps and identify emerging evaluation practices.

H. Limitations

Several limitations should be acknowledged. First, the review covered the period from January 2019 to July 2025 and focused on peer-reviewed English-language literature, excluding grey literature and non-English sources. Second, the quality assessment involved a degree of researcher judgment, particularly in borderline cases. Third, while metric types and calculation methods were standardised to support comparison, some hybrid approaches may not have been fully captured by this classification. Lastly, the inclusion criteria required that transparency be explicitly evaluated, which may have led to the exclusion of relevant studies where transparency was subsumed under broader constructs such as fairness or trustworthiness.

V. Results

This section presents the results of the systematic literature review, providing a descriptive overview of the selected studies, including the classification of evaluation methods, the distribution of metric types and calculation approaches, as well as the sectoral focus, implementation status, and references to relevant legal and ethical frameworks.

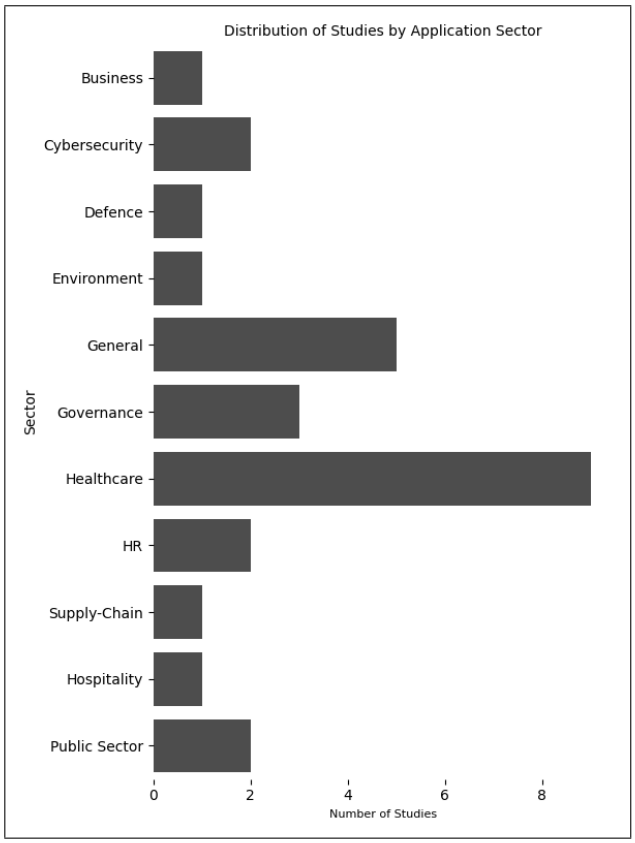

Fig. 3. Sector distribution of the 28 included studies; healthcare 9/28 (32%).

A. Overview of Included Studies

A total of 28 peer-reviewed studies published between January 2019 and July 2025 were selected for inclusion. The majority introduced formal frameworks or tools aimed at evaluating transparency in artificial intelligence systems, while a smaller subset proposed conceptual approaches grounded in governance or policy-oriented perspectives. In terms of application status, most of the studies (23 out of 28) applied their method to a real-world system, dataset, or use case. At the same time, the rest remained conceptual, theoretical, or exploratory in scope. See Fig. 2 for the year-by-year distribution.

The analysis revealed a notable concentration of transparency evaluation in the healthcare sector, with 9 of the 28 studies (32%) focusing on clinical decision support, medical imaging, or digital health systems [19], [20], [21], [4], [22], [23], [24], [25], [26]. Other studies proposed general-purpose frameworks or were situated in domains such as cybersecurity, HR, or public sector applications. A smaller number of studies addressed domains such as hospitality or defence. This sectoral imbalance highlights the strong regulatory and ethical pressures in healthcare, in contrast to more limited methodological development in other fields. See Fig. 3 for sectoral concentration.

B. Metric Types

The studies were categorised according to the type of metric or evaluation logic employed. Seven primary metric

types were identified:

- Checklist-based approaches operationalise transparency through binary questions or structured assessment lists. [27][19][28][25][29][30][17][31][20]
- Qualitative methods rely on expert judgement or narrative analysis. [21][4][32][33][34]
- Explainability-based evaluations assess transparency through interpretability outputs (e.g., SHAP, LIME). [35][36][37][38]
- Survey-based tools capture perceptions of transparency using structured questionnaires. [39][22][40][41]
- Index-based approaches produce composite scores or traceability indexes. [26][24][42]
- Score-based methods generate transparency scores by assigning weighted values to defined criteria. [43][23]
- Rating-based frameworks. [44]

Checklist-based approaches were the most common, often reflecting broader trends in AI governance tools such as the ALTAI framework [17].

C. Regulatory Alignment

References to legal and ethical frameworks were present in 19 of the 28 studies (67%) [19], [17], [31], [20], [4], [39], [22], [44], [33], [27], [34], [36], [42], [28], [24], [25], [38], [26], [29], most commonly the EU Artificial Intelligence Act, the General Data Protection Regulation (GDPR), and the AI HLEG Guidelines. However, only a few studies provided traceable mappings between evaluation criteria and specific legal provisions. For instance, RETENTION [26] and POLARIS [28] embedded elements of the EU AI Act [2] and AI HLEG recommendations [17] within their structure. In the majority of cases, alignment remained high-level or aspirational, with limited support for formal compliance, certification, or auditability.

D. Quality Assessment

Across the 28 studies, the overall quality was high, with most achieving at least 70% of the maximum possible score (see Table II). Several studies attained full or nearly full scores, while others performed lower due to conceptual or preliminary scope.

Specifically, QCA1 (evaluation method outlined) scored the highest, achieving 96.4% of the maximum score. This outcome was expected, as the presence of an explicitly described transparency evaluation method was a prerequisite for inclusion during the full-text review. QCA3 (discussion of strengths and limitations) also performed strongly, with 94.6%. By contrast, QCA2 (specific criteria for transparency assessment) and QCA5 (use of real-world AI applications) achieved 76.8% each, indicating room for improvement in applying structured criteria and real-world validation. The lowest-performing criterion was QCA4 (comparison with other transparency frameworks), with only 46.4% of the maximum score, showing that relatively few studies benchmarked their approach against other methods.

As shown in Table II, studies with higher total scores were typically more detailed, provided stronger methodological justification, and included real-world applications. In contrast, lower-scoring studies tended to be more conceptual or exploratory. These quality assessment results guided the interpretation of the final analysis, following best practices from prior systematic reviews (e.g., [14]), and support a nuanced discussion of strengths and limitations reported in Section Thematic Analysis.

VI. Thematic Analysis

This section presents the thematic analysis of the 28 studies included in the review. The analysis identified six overarching themes that characterise how transparency is defined, operationalised, and evaluated in artificial intelligence systems. These themes reflect both recurring strengths and persistent limitations across the literature.

A. Theme 1: Transparency as a Multi-Dimensional Construct

A consistent finding across the reviewed studies is the conceptualisation of transparency as a composite of multiple, interrelated dimensions. This mirrors the influential framing of the High-Level Expert Group on AI, which identified transparency, encompassing traceability, explainability, and communication, as one of the seven key requirements for Trustworthy AI [1]. These three pillars provided an early normative anchor and were frequently echoed, either explicitly or implicitly, in the SLR dataset.

Traceability concerns the documentation, logging, and oversight of system development, data flows, and decision-making processes, enabling stakeholders to follow and verify AI behaviour across its lifecycle. Explainability focuses on the interpretability of models or outputs, employing techniques such as SHAP, LIME, or saliency maps to make system reasoning intelligible. Communication refers to the disclosure of relevant information to users, regulators, or affected individuals in ways that are accessible, usable, and contextually appropriate.

While most studies referenced all three pillars, few operationalised them in an integrated manner. For example, IntelliLung [24], RETENTION [26], and SMACTR [30] explicitly combined traceability mechanisms, explainability techniques, and user-facing communication strategies within a single evaluation framework. In contrast, most approaches prioritised one or two dimensions, with communication often underdeveloped or absent. This imbalance reflects a gap in translating high-level principles into holistic evaluation designs.

This multi-dimensional perspective aligns closely with the stakeholder-centred framework of Ofem et al. [16], where

TABLE II. QUALITY ASSESSMENT RESULTS FOR INCLUDED STUDIES (N=28)

Study ID	QCA1	QCA2	QCA3	QCA4	QCA5	Total	% of Max
PS1	1	1	1	0.5	1	4.5	90%
PS2	1	0.5	0.5	0.5	0	2.5	50%
PS3	1	0.5	1	0.5	1	4.0	80%
PS4	1	0.5	1	0	1	3.5	70%
PS5	1	1	0.5	0	0	2.5	50%
PS6	1	0.5	1	0	0.5	3.0	60%
PS7	1	1	1	1	0	4.0	80%
PS8	1	0.5	1	0.5	1	4.0	80%
PS9	1	0.5	1	0.5	1	4.0	80%
PS10	1	0.5	1	0	1	3.5	70%
PS11	1	1	1	0.5	1	4.5	90%
PS12	1	1	1	1	1	5.0	100%
PS13	0.5	0.5	1	1	0	3.0	60%
PS14	1	1	0.5	0.5	1	4.0	80%
PS15	1	1	1	0.5	1	4.5	90%
PS16	1	1	1	0.5	1	4.5	90%
PS17	0.5	0.5	1	0	0.5	2.5	50%
PS18	1	1	1	0.5	1	4.5	90%
PS19	1	0.5	1	0.5	1	4.0	80%
PS20	1	1	1	0.5	1	4.5	90%
PS21	1	0.5	1	0.5	1	4.0	80%
PS22	1	1	1	0.5	0	3.5	70%
PS23	1	0.5	1	1	1	4.5	90%
PS24	1	1	1	0.5	1	4.5	90%
PS25	1	0.5	1	0	1	3.5	70%
PS26	1	1	1	0	0.5	3.5	70%
PS27	1	1	1	1	1	5.0	100%
PS28	1	1	1	0.5	1	4.5	90%
Total	27	21.5	26.5	13	21.5	109.5	—
% of Max	96.4%	76.8%	94.6%	46.4%	76.8%	—	—

transparency is linked to quality factors such as accessibility, usability, and understandability. Mapping the three pillars to these factors revealed that traceability overlaps most strongly with accessibility and modifiability, explainability with understandability, and communication with accessibility and usability. This mapping provides a bridge between AI ethics discourse and measurable software engineering practices.

In the context of this SLR, the prevalence of these three pillars directly addresses RQ1 (existing approaches to transparency evaluation) by highlighting their role as foundational building blocks. However, their uneven treatment across studies also informs RQ3 (strengths, limitations, and gaps), underscoring the need for integrated approaches that treat traceability, explainability, and communication as mutually reinforcing rather than isolated components.

B. Theme 2: Dominance of Checklists and Qualitative Methods

Across the 28 included studies, checklist-based approaches were the single most common way to evaluate transparency (*9/28; 32%*) [27], [19], [28], [25], [29], [30], [17], [31], [20], followed by scenario-based qualitative assessments (*5/28; 18%*) [21], [4], [32], [33], [34]. Explainability-output-driven evaluations (e.g., using SHAP/LIME artefacts) and survey instruments each appeared in *4/28* studies (*14%* apiece), while index-style composites were less frequent (*3/28; 11%*). Score-style metrics were rare (*2/28; 7%*), and explicit rating-scale schemes appeared only once (*1/28; 4%*).[1]

This pattern is consistent with the field's reliance on governance-aligned checklists (e.g., ALTAI-inspired instruments) and structured internal-audit playbooks (e.g., SMACTR; POLARIS; MLOps audit catalogues), which are accessible and broadly usable across domains [17], [30], [28], [25]. However, checklist implementations showed variable methodological quality in our review: while many were robust, some scored lower on our quality assessment (e.g., overall study QA scores in the 2.5-4.5 range), and, critically, few offered granular scoring, maturity levels, or CI/CD-friendly automation. In short, they structure the conversation, but they rarely quantify transparency in a way that is reproducible or benchmarkable across settings.

Significant counterexamples highlight how quantification can evolve. The Visibility Index Score (VIS) adapts supply-chain "visibility" logic to AI pipelines, scoring documentation quantity, freshness, and accuracy [43]. The Cyrus framework automates parts of transparency evaluation via FAIR-aligned metadata and linked-data quality tooling [23]. Both illustrate promising routes toward measurable, repeatable assessments—yet each remains bounded by domain and artefact conventions (VIS by reviewer judgment; Cyrus by metadata-rich, RDF/FAIR contexts). These limits inform our subsequent design choices: transitioning from static checklists to artefact-level metadata, evaluation linkages, and scorecard-driven gap closure that remain lightweight enough for real-world projects.

[1]These tallies refer to the dominant metric type per study when hybrids were present.

C. Theme 3: Limited Stakeholder and User Involvement

A striking gap in the reviewed literature is the limited inclusion of stakeholders, especially end-users, in the design, execution, and validation of transparency evaluations. Only a minority of studies (*6/28; 21%*) explicitly incorporated stakeholder perspectives into their methods, typically via structured interviews, surveys, or participatory workshops. Examples include IntelliLung [24], which involved clinicians to identify the most relevant artefacts and interpretability measures in a clinical decision-support context, and the ALTAI-inspired sectoral adaptations [27], [25], which consulted domain experts to weight checklist items. However, these engagements were generally one-off and did not establish feedback loops to iteratively refine the evaluation over the AI system lifecycle.

Despite regulatory emphasis on transparency being understandable to affected persons [2], only a small number of studies incorporate user perspectives into their evaluation frameworks. FAT-CAT [40] and the survey-based method by Fehr et al. [22] are exceptions, explicitly including perceptions of transparency, trust, and understanding in their measurement schemes.

A significant outlier to the lack of continuous stakeholder engagement is the Z-Inspection® Framework [4], which embeds a structured feedback loop across its Set-Up, Assess, and Resolve phases. While the stakeholders involved are often domain experts, project partners, or regulators rather than the ultimate end-users, the framework ensures that transparency requirements are revisited and refined after scenario analysis and ethical issue mapping, and that unresolved gaps feed into improvement actions. This iterative design demonstrates how multi-phase engagement can be operationalised, even if end-user representation remains limited.

From a software engineering perspective, the omission of such feedback mechanisms in most studies conflicts with stakeholder-driven models such as Ofem et al.'s transparency framework [16], which positions stakeholder requirements as the starting point for defining quality factors and associated metrics. Without ongoing stakeholder input, whether from end-users or intermediary stakeholders, transparency evaluations risk being misaligned with the information needs, literacy levels, and contextual constraints of those most affected by AI systems.

In relation to our research questions, this theme directly addresses RQ2 (how key transparency concerns are addressed) by showing that while stakeholder relevance is widely recognised in principle, it is rarely operationalised in a structured and continuous manner. It also informs RQ3 (strengths, limitations, and gaps), identifying the lack of sustained, multi-phase stakeholder participation as a significant limitation in current transparency evaluation practice.

D. Theme 4: Sectoral Concentration in Healthcare

Healthcare emerged as the dominant application domain for transparency evaluation frameworks in our SLR, with **9/28** studies (**32%**) focusing exclusively on medical AI systems. These ranged from diagnostic imaging and clinical decision-support tools to patient-facing health applications[19], [20], [21], [4], [22], [23], [24], [25], [26].

The concentration reflects healthcare's high regulatory scrutiny, established ethical oversight mechanisms, and the availability of structured artefacts such as clinical guidelines, audit logs, and model validation reports.

Relevant examples include IntelliLung [24], which applied an integrated traceability, explainability, and communication framework to assess an AI system for lung disease diagnosis, and RETENTION [26], which evaluated an AI tool for predicting patient readmission risks. Both demonstrated the feasibility of combining multiple transparency pillars with sector-specific artefacts and metrics, benefiting from healthcare's pre-existing documentation and audit cultures.

In contrast, sectors such as finance, transportation, and public administration were underrepresented, with most non-healthcare cases relying on adapted checklists (e.g., ALTAI) or scenario-based ethical assessments. These domains often lack the same level of artefact maturity or regulatory compulsion, resulting in evaluations that are either less comprehensive or highly context-dependent.

This imbalance has implications for RQ4 (applicability in real-world AI systems). While healthcare-focused frameworks offer valuable methodological depth, their artefact-rich environment may not translate directly to sectors where documentation practices are less formalised. For cross-sector applicability, transparency evaluation methods must adapt to different artefact availability, stakeholder profiles, and regulatory drivers without sacrificing rigour. This supports the need, identified in RQ3, for domain-flexible frameworks that can scale from artefact-rich to artefact-sparse environments.

E. Theme 5: Partial and Informal Regulatory Alignment

Many transparency evaluation approaches in the reviewed studies reference regulatory or ethical frameworks, most commonly the EU AI Act, the AI HLEG Ethics Guidelines for Trustworthy AI, or sector-specific standards. However, in *20/28* studies (*71%*), this alignment was *partial* or *informal*: regulatory documents were cited as conceptual anchors, but their specific obligations were not systematically translated into evaluation criteria or metrics.

For example, several healthcare-focused studies mapped their checklists to the AI HLEG requirements [17] or to medical device regulations, but did not provide a traceable mapping between regulatory clauses (e.g., EU AI Act Article 13 obligations for high-risk systems) and the artefacts or scores used in their evaluations. Similarly, frameworks such as ALTAI were adapted to specific organisational contexts without maintaining the whole structure or weighting defined in the original.

Only a small subset of studies (*4/28; 14%*) achieved strong formal alignment by creating explicit mappings from regulatory text to artefacts, evaluation metrics, and scoring logic. Examples include RETENTION [26] and POLARIS [28]. These approaches demonstrated clearer audit readiness and traceability of compliance claims but required significant up-front investment in legal-technical translation.

In relation to our research questions, this theme informs both RQ2 (how key transparency concerns are addressed) and RQ3 (strengths, limitations, and gaps). It shows that while

regulatory references are widespread, operationalisation into concrete, verifiable measures is inconsistent. This gap highlights the importance of evaluation frameworks that integrate regulatory requirements directly into artefact-metric mappings, allowing for compliance checks alongside broader ethical evaluations.

F. Theme 6: Gaps in Standardisation and Lifecycle Integration

A recurring limitation across the reviewed studies is the lack of standardisation in how transparency is defined, measured, and maintained over an AI system's lifecycle. Even when similar terminology is used (e.g., "traceability" or "model interpretability"), the underlying definitions, artefacts, and metrics vary considerably, making cross-study comparison and benchmarking difficult. This variability extends to the granularity of evaluation (system-level vs. component-level), the form of evidence accepted (qualitative descriptions vs. quantitative scores), and the frequency of assessment.

Lifecycle integration is also underdeveloped. Most frameworks focus on a single point-in-time evaluation, often predeployment or post-deployment, without mechanisms for reassessment as the system evolves. Continuous monitoring, when mentioned, was generally conceptual rather than operationalised through automated pipelines or version-controlled artefact registries.

A few exceptions, such as Cyrus [23], VIS [43], and Z-Inspection® [4], incorporate repeatable measurement processes and artefact tracking. In particular, Z-Inspection® embeds a structured feedback loop across its Set-Up, Assess, and Resolve phases, enabling transparency requirements to be revisited and refined in light of new evidence or system changes. While the involved stakeholders are typically domain experts or project partners rather than end-users, this lifecycle-aware design demonstrates how continuous evaluation can be operationalised in practice.

A number of studies also reported unintended consequences of transparency practices. For instance, Lee et al. [40] noted increased user anxiety when information was disclosed without sufficient context, while Fehr et al. [22] highlighted concerns from developers regarding the exposure of proprietary information. These findings illustrate the need to balance transparency with usability, commercial confidentiality, and cognitive burden, ensuring that lifecycle-integrated evaluations do not inadvertently reduce user trust or hinder adoption.

This theme directly addresses RQ3 (strengths, limitations, and gaps) by identifying the absence of standardised definitions and cross-domain benchmarks, and RQ4 (applicability in real-world AI contexts) by showing that without lifecycle integration, transparency evaluation risks becoming a compliance snapshot rather than an ongoing governance practice. The findings support the need for artefact-driven, metadata-based approaches that can be embedded into development workflows, enabling continuous, comparable, and stakeholder-relevant transparency evaluation.

VII. Discussion

This section synthesises the key findings of the review in relation to the four research questions. It reflects on the strengths and limitations of current transparency evaluation practices in artificial intelligence, highlights existing gaps, and considers implications for future research and practice.

A. RQ1: Existing Approaches to Transparency Evaluation

Many studies introduced formal instruments such as ALTAI [17], POLARIS [28], and RETENTION [26], which aim to structure transparency evaluation through checklists or composite scoring. Others, like Z-Inspection® [4], offer qualitative, scenario-based frameworks grounded in ethical deliberation and stakeholder reflection. While these approaches enhance contextual sensitivity, they do not offer quantitative scoring or automation features.

Collectively, they reflect a growing recognition of the need to move beyond abstract principles and towards operational methods. Despite this progress, a clear pattern emerged: the majority of studies adopt qualitative or semi-structured methods, with limited attention to scoring granularity, automation, or lifecycle integration. Checklist-based approaches remain dominant but are often used without weighting, maturity models, or replicable metrics. Few studies propose evaluation instruments that can be implemented at scale or embedded into technical workflows.

B. RQ2: How Key Transparency Concerns are Addressed

Thematic analysis indicates that transparency is predominantly conceptualised as a multi-dimensional construct encompassing traceability, explainability, and communication. While most studies address at least one of these pillars, only a limited number propose integrated frameworks that operationalise all three in a cohesive manner.

Traceability and explainability are the most frequently addressed dimensions, commonly through system documentation, model-level interpretation methods, or audit trails. In contrast, communication, especially toward non-technical users or impacted stakeholders, remains significantly underdeveloped across the literature.

Although many frameworks reference values such as trust, usability, or ethical alignment, only a minority explicitly incorporate stakeholder or end-user perspectives in the evaluation process. Notable exceptions include the FAT-CAT framework [40] and the transparency survey developed by Fehr et al. [22], both of which engage user feedback more directly. However, such examples remain rare.

This highlights a persistent gap between the normative emphasis on stakeholder-centred transparency and the current reality of system- or expert-centric evaluation practices. Most frameworks still prioritise internal validation or technical explainability over participatory or communicative transparency mechanisms.

C. RQ3: Strengths, Limitations, and Gaps

Several strengths can be observed in the reviewed literature. First, there is increasing convergence around the idea that transparency must be evaluable, not merely aspirational. Second, the inclusion of sector-specific indicators, particularly in healthcare, demonstrates the potential for contextual adaptation. Third, some studies have made significant strides

in mapping transparency to legal or ethical requirements. Nonetheless, persistent limitations were identified:

- Over-reliance on checklists and qualitative assessments, with limited support for scoring, automation, or benchmarking.
- Lack of standardised frameworks or maturity models for transparency.
- Limited stakeholder inclusion, especially in the evaluation of communication clarity or user understanding.
- Inconsistent treatment of lifecycle phases, with transparency often evaluated only at the design stage or after.
- Conceptual conflation between transparency, explainability, and related constructs.

These limitations mirror concerns raised in prior literature. For instance, Mittelstadt et al. [45] warned against ambiguity in transparency definitions, arguing that a lack of clarity undermines both the interpretability of AI systems and the accountability of their developers.

D. RQ4: Applicability in Real-World AI Contexts

Of the 28 studies reviewed, 23 (82%) [43], [30], [41], [31], [21], [4], [39], [22], [23], [44], [40], [35], [33], [27], [34], [36], [37], [28], [24], [25], [38], [26], [29] were applied in real-world contexts, suggesting a growing emphasis on empirical validation. These included pilot implementations, case studies, and evaluations of deployed systems, particularly in healthcare and cybersecurity domains. However, even among applied studies, evaluation methods often lacked formal mechanisms for compliance verification, stakeholder feedback loops, or documentation updates over time. Moreover, the strong concentration of studies in healthcare raises questions about generalisability. While medical AI offers a high-risk, high-regulation testbed, other domains, such as education, employment, or defence, are underexplored. This suggests that many existing frameworks may not yet be transferable to broader or less-regulated contexts.

E. Additional Observations and Future Trends

Emerging findings suggest that increased transparency can paradoxically raise user anxiety or cognitive burden, as demonstrated by Lee et al. [40]. Furthermore, several studies highlight the potential need for transparency toolkits, particularly for SMEs, to operationalise legal requirements in accessible ways. Concepts such as real-time AI dashboards ('AI cockpits') [33] or safe harbour mechanisms [31] also represent future directions worth exploring.

F. Synthesis and Implications for Future Research

Taken together, the findings highlight both progress and persistent gaps in the operationalisation of transparency in AI systems. While concrete evaluation methods are emerging, their diversity in scope, terminology, and measurement approaches underscores a lack of standardisation that complicates benchmarking across domains.

1) Fragmentation and standardisation challenges: Similar concepts, such as traceability or interpretability, are often operationalised differently, with artefacts ranging from checklists and surveys to explainability-based metrics. This fragmentation limits cross-domain comparability and reinforces the need for adaptable yet standardised schemes.

2) Lifecycle integration and sustainability: Most methods assess transparency only at discrete points, making it a compliance "snapshot" rather than an ongoing governance practice. Exceptions such as Cyrus and VIS point to lifecycle-aware approaches, but these remain isolated and not widely adopted.

3) Stakeholder perspectives and unintended consequences: Few studies systematically include user perspectives. Where they do, such as FAT-CAT [40] or Fehr et al. [22], results show both benefits and unintended consequences (e.g., increased anxiety when disclosure lacks context, or concerns about exposing proprietary information). This underscores the tension between usability, confidentiality, and cognitive burden.

4) Quality assessment insights: The quality assessment (QCA1-QCA5) provides additional perspective. While QCA1 (clarity of evaluation method) and QCA3 (discussion of strengths and limitations) scored highly (over 90%), QCA4 (comparison with other frameworks) was much weaker at 46.4%. As a result, transparency evaluation remains fragmented and difficult to contextualise, undermining comparability across sectors.

5) Bridging ethics and engineering: The three-pillar framing of traceability, explainability, and communication, first articulated by the AI HLEG [1], offers a conceptual anchor. Mapping these pillars to quality factors from software engineering (e.g., accessibility, usability, understandability) suggests a bridge between high-level ethical principles and measurable lifecycle practices.

Beyond these priorities, the review highlights an overarching limitation: the lack of artefact-driven methods capable of embedding transparency evaluation into day-to-day development workflows. Existing frameworks rarely provide mechanisms for capturing, standardising, and reusing transparency artefacts across system versions and contexts. This is a critical gap given the need for reproducibility, comparability, and regulatory accountability in real-world AI applications.

Addressing this limitation directly informed our subsequent work—a companion paper, which extends the Z-Inspection® process within the MANOLO context by proposing a Transparency Artefact Registry. TAR operationalises the insights of this SLR by offering a structured, metadata-based approach for cataloguing transparency artefacts, linking them to stakeholder requirements, and supporting continuous lifecycle evaluation. In doing so, it aims to provide researchers, practitioners, and regulators with a verifiable and repeatable mechanism for demonstrating transparency in line with European regulatory and ethical standards.

6) Design implications: The deficits identified map to an artefact-driven remediation. Missing standardisation calls

for a minimal metadata schema that normalises transparency artefacts across projects. Weak lifecycle integration requires a versioned registry that tracks artefacts across releases and updates. Informal regulatory alignment is addressed by explicit field-level mappings from artefacts to EU AI Act obligations (e.g., Article 13 on traceability and user information). Limited stakeholder inclusion is realised by tagging artefacts with stakeholder roles and providing communication-ready views. These elements motivate the TAR and align with Z-Inspection®.

Ultimately, transparency should not be treated as a static artefact or the outcome of a checklist, but as a dynamic, context-sensitive, and stakeholder-driven attribute that must be continuously embedded throughout the AI system development lifecycle.

VIII. Conclusion

This paper has presented an SLR of 28 peer-reviewed studies on transparency evaluation in artificial intelligence systems. The analysis revealed that while transparency is widely recognised as a requirement for trustworthy and accountable AI, current evaluation methods remain fragmented. Most approaches rely on qualitative checklists or scenario-based assessments, with limited support for scoring, standardisation, or lifecycle integration. Stakeholder involvement, particularly from end-users, is still rare. In particular, the communication dimension of transparency, meaning the disclosure and presentation of information about AI systems to users, regulators, and other stakeholders in accessible and understandable ways, remains underdeveloped compared to traceability and explainability.

Despite these limitations, the review identified encouraging trends. Several studies successfully link transparency evaluation to regulatory or ethical requirements, others adapt sector-specific indicators (notably in healthcare), and a minority demonstrate mechanisms for empirical validation in real-world systems. Together, these works mark a gradual shift from aspirational principles to operational practices.

At the same time, persistent gaps were observed: the lack of shared taxonomies and maturity models, insufficient mechanisms for continuous evaluation across the AI lifecycle, and the absence of artefact-driven methods for capturing and reusing transparency evidence. These findings highlight both the strengths of current approaches and the need for more systematic, repeatable, and scalable solutions.

The results of this review directly inform our subsequent work within the context of the MANOLO project, where we extend the Z-Inspection® Framework with a Transparency Artefact Registry (TAR). By cataloguing transparency artefacts through structured metadata and linking them to stakeholder requirements and regulatory obligations, TAR provides a practical means of addressing the limitations identified in this review. In this way, the SLR serves as both a state-of-the-art assessment and the conceptual foundation for developing artefact-based mechanisms that enable transparency to be continuously evaluated and demonstrably aligned with the EU AI Act.

In sum, transparency evaluation is moving from principle to practice but remains methodologically fragmented and weakly grounded across the AI lifecycle. Closing this gap requires standardised taxonomies, comparable metrics, and evidence-bearing artefacts embedded in development and compliance workflows.

Future research should therefore focus on refining such artefact-driven approaches, testing their applicability across domains beyond healthcare, and ensuring that stakeholder perspectives are integrated in a sustained and iterative manner. Only by embedding transparency throughout the lifecycle of AI systems can it become a verifiable attribute of trustworthy AI rather than a static aspiration.

Acknowledgment

This work is based on and extends the thesis by Karanxha [46], completed at Laurea University of Applied Sciences, Finland. This work was co-funded by the European Union under GA no. 101135782. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or CNECT. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] High-Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy ai," European Commission, 2019, accessed 17 May 2025. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[2] "Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act)," Official Journal of the European Union, L 168/1, 2024, accessed 17 May 2025. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689

[3] "Manolo project," https://manolo-project.eu/, 2023, accessed: 17-Aug-2025.

[4] R. V. Zicari, J. Brusseau, S. N. Blomberg, P. J. Honkoop, J. Haverinen, C. Kuziemsky, J. Laaksolahti, A. Lang, P. Linko, S. Mahmood, and et al., "Z-inspection®: A process to assess trustworthy ai," *IEEE Transactions on Technology and Society*, vol. 2, no. 2, pp. 83–97, 2021. [Online]. Available: https://doi.org/10.1109/TTS.2021.3066209

[5] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, "Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai," Berkman Klein Center for Internet & Society, Harvard University, Research Publication 2020-1, 2020. [Online]. Available: https://dx.doi.org/10.2139/ssrn.3518482

[6] M. Ebers, V. R. S. Hoch, F. Rosenkranz, H. Ruschemeier, and B. Steinrötter, "The european commission's proposal for an artificial intelligence act—a critical assessment by members of the robotics and ai law society (rails)," *J*, vol. 4, no. 4, pp. 589–603, 2021. [Online]. Available: https://doi.org/10.3390/j4040043

[7] R. A. Agha, G. Mathew, R. Rashid, A. Kerwan, A. Al-Jabir, C. Sohrabi, T. Franchi, M. Nicola, M. Agha, and T. Group, "Transparency in the reporting of artificial intelligence – the titan guideline," *Premier Journal of Science*, vol. 10, p. 100082, 2025. [Online]. Available: https://doi.org/10.70389/PJS.100082

[8] S. Genovesi, M. Haimerl, I. Merget, S. M. Prange, O. Obert, S. Wolf, and J. Ziehn, "Evaluating dimensions of ai transparency: A comparative study of standards, guidelines, and the eu ai act," in *Symposium on Scaling AI Assessments (SAIA 2024)*, ser. OpenAccess Series in Informatics (OASIcs), vol. TBD. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2025, pp. 10:1–10:17. [Online]. Available: https://doi.org/10.4230/OASIcs.SAIA.2024.10

[9] M. Busuioc, "Accountable artificial intelligence: Holding algorithms to account," *Public Administration Review*, vol. 81, no. 5, pp. 825–836, 2021, open Access. [Online]. Available: https://doi.org/10.1111/puar.13293

[10] C. Panigutti, R. Hamon, I. Hupont, D. Fernandez Llorca, D. Fano Yela, H. Junklewitz, S. Scalzo, G. Mazzini, I. Sanchez, J. Soler Garrido, and E. Gomez, "The role of explainable ai in the context of the ai act," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2023, pp. 1139–1150. [Online]. Available: https://doi.org/10.1145/3593013.3594069

[11] J. C. S. d. P. Leite and C. Cappelli, "Software transparency," *Business & Information Systems Engineering*, vol. 2, no. 3, pp. 127–139, 2010.

[12] M. Hosseini, A. Shahri, K. Phalp, and R. Ali, "Engineering transparency requirements: A modelling and analysis framework," *Information Systems*, vol. 74, pp. 3–22, 2018, information Systems Engineering: selected papers from CAiSE 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306437916305282

[13] D. Spagnuelo, C. Bartolini, and G. Lenzini, "Qualifying and measuring transparency: A medical data system case study," *Computers & Security*, vol. 91, p. 101717, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016740481830823X

[14] P. Ofem, B. Isong, and F. Lugayizi, "On the concept of transparency: A systematic literature review," *IEEE Access*, vol. 10, pp. 89 887–89 914, 2022. [Online]. Available: https://doi.org/10.1109/ACCESS.2022.3200487

[15] B. Isong, P. Ofem, and F. Lugayizi, "Towards a framework for improving transparency in the software engineering process," in *Proceedings of the 2022 12th International Conference on Software Technology and Engineering (ICSTE)*, Osaka, Japan, 2022, pp. 19–28. [Online]. Available: https://doi.org/10.1109/ICSTE57415.2022.00011

[16] P. Ofem, B. Isong, and F. Lugayizi, "Metrics for evaluating and improving transparency in software engineering: An empirical study and improvement model," *SN Computer Science*, vol. 5, no. 1097, pp. 1–16, 2024.

[17] High-Level Expert Group on Artificial Intelligence, "The assessment list for trustworthy artificial intelligence (altai) for self-assessment," Report, Brussels, 2020, accessed 17 May 2025. [Online]. Available: https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

[18] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006. [Online]. Available: https://doi.org/10.1191/1478088706qp063oa

[19] R. Rokhshad, M. Ducret, A. Chaurasia, T. Karteva, M. Radenkovic, J. Roganovic, M. Hamdan, H. Mohammad-Rahimi, J. Krois, P. Lahoud, and F. Schwendicke, "Ethical considerations on artificial intelligence in dentistry: A framework and checklist," *Journal of Dentistry*, 2023.

[20] M. Mora-Cantallops, S. Sánchez-Alonso, E. García-Barriocanal, and M.-A. Sicilia, "Traceability for trustworthy ai: A review of models and tools," *Big Data and Cognitive Computing*, vol. 5, no. 2, p. 20, 2021. [Online]. Available: https://doi.org/10.3390/bdcc5

[21] D. W. Joyce, A. Kormilitzin, K. A. Smith, A. Simons, I. Craddock, and A. Nevado-Holgado, "Explainable artificial intelligence for mental health through transparency and interpretability for understandability," *npj Digital Medicine*, vol. 6, p. 6, 2023. [Online]. Available: https://doi.org/10.1038/s41746-023-00751-9

[22] J. Fehr, G. Jaramillo-Gutierrez, L. Oala, M. I. Gröschel, M. Bierwirth, P. Balachandran, A. Werneck-Leite, and C. Lippert, "Piloting a survey-based assessment of transparency and trustworthiness with three medical ai tools," *Healthcare*, vol. 10, no. 10, p. 1923, 2022. [Online]. Available: https://doi.org/10.3390/healthcare10101923

[23] M. Basereh, A. Caputo, and R. Brennan, "Automatic transparency evaluation for open knowledge extraction systems," *Journal of Biomedical Semantics*, vol. 14, p. 12, 2023. [Online]. Available: https://doi.org/10.1186/s13326-023-00293-9

[24] V. Janev, M. Nenadović, D. Paunović, S. Vahdati, J. Li, M. H. Yousuf, J. Montanya, R. Theilen, J. Wittenstein, S. Tsurkan, and R. Huhle, "IntelliLung AI-DSS Trustworthiness Evaluation Framework," in *Proceedings of the 2024 32nd Telecommunications Forum (TELFOR)*. IEEE, 2024. [Online]. Available: https://doi.org/10.1109/TELFOR63250.2024.10819068

[25] L. Helmer, C. Martens, D. Wegener, M. Akila, D. Becker, and S. Abbas, "Towards trustworthy ai engineering: A case study on integrating an ai audit catalog into mlops processes," in *Proceedings of the 2nd International Workshop on Responsible AI Engineering (RAIE '24)*. New York, NY, USA: Association for Computing Machinery, 2024, pp. 1–7. [Online]. Available: https://doi.org/10.1145/3643691.3648584

[26] I. E. Nicolae, G. Danciu, C. Nanou, N. Koulierakis, and V. Danilatou, "Transparency metrics for artificial intelligence-driven applications in healthcare," in *Proceedings of the 13th Hellenic Conference on Artificial Intelligence (SETN '24)*. New York, NY, USA: Association for Computing Machinery, 2024, pp. 1–8. [Online]. Available: https://doi.org/10.1145/3688671.3688782

[27] D. M. F. Saldanha, C. N. Dias, and S. Guillaumon, "Transparency and accountability in digital public services: Learning from the brazilian cases," *Government Information Quarterly*, vol. 39, no. 2, p. 101680, 2022. [Online]. Available: https://doi.org/10.1016/j.giq.2022.101680

[28] M. T. Baldassarre, D. Gigante, M. Kalinowski, and A. Ragone, "POLARIS: A framework to guide the development of trustworthy AI systems," in *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI (CAIN '24)*. New York: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3644815.3644947

[29] P. Salehi, Y. Ba, N. Kim, A. Mosallanezhad, A. Pan, M. C. Cohen, Y. Wang, J. Zhao, S. Bhatti, J. Sung, E. Blasch, M. V. Mancenido, and E. K. Chiou, "Towards trustworthy ai-enabled decision support systems: Validation of the multisource ai scorecard table (mast)," *Journal of Artificial Intelligence Research*, 2024. [Online]. Available: https://doi.org/10.1613/jair.1.14990

[30] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, 2020, pp. 33–44.

[31] S. Yanisky-Ravid and S. K. Hallisey, ""Equality and Privacy by Design"": A New Model of Artificial Intelligence Data Transparency via Auditing, Certification, and Safe Harbor Regimes," *Fordham Urban Law Journal*, vol. 46, no. 2, pp. 428–491, 2019, accessed 17 May 2025. [Online]. Available: https://ir.lawnet.fordham.edu/ulj/vol46/iss2/5

[32] N. A. Sharma, R. R. Chand, Z. Buksh, A. B. M. S. Ali, A. Hanif, and A. Beheshti, "Explainable ai frameworks: Navigating the present challenges and unveiling innovative applications," *Algorithms*, vol. 17, no. 6, p. 227, 2024. [Online]. Available: https://doi.org/10.3390/a17060227

[33] L. Jantzen, M. Bottel, and R. Kempen, "How to achieve trust, satisfaction, and acceptance in the interaction with ai through an ai cockpit and a transparency interface? – a psychological framework," *Procedia Computer Science*, vol. 246, pp. 292–301, 2024. [Online]. Available: https://doi.org/10.1016/j.procs.2024.09.408

[34] A. Dubey, Z. Yang, and G. Hattab, "A nested model for ai design and validation," *iScience*, vol. 27, no. 9, p. 1110603, 2024, open Access. [Online]. Available: https://doi.org/10.1016/j.isci.2024.1110603

[35] K. Sokol, R. Santos-Rodriguez, and P. Flach, "Fat forensics: A python toolbox for algorithmic fairness, accountability and transparency," *Software Impacts*, vol. 14, p. 100406, 2022. [Online]. Available: https://doi.org/10.1016/j.simpa.2022.100406

[36] J. Haurogné, N. Basheer, and S. Islam, "Vulnerability detection using bert based llm model with transparency obligation practice towards trustworthy ai," *Machine Learning with Applications*, vol. 18, p. 100598, 2024. [Online]. Available: https://doi.org/10.1016/j.mlwa.2024.100598

[37] S. V. S. Kumar and H. K. Kondaveeti, "Towards transparency in ai: Explainable bird species image classification for ecological research," *Ecological Indicators*, vol. 169, p. 112886, 2024. [Online]. Available: https://doi.org/10.1016/j.ecolind.2024.112886

[38] A. Bansal, D. A. A. K, S. Gangadharan, R. R. R. R, M. M. Ismail, and N. K. A. Halaf, "Ethical considerations of AI implementation in business planning: Ensuring fairness and transparency," in *Proceedings of the 2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, Chennai, India, 2023, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ICSES60034.2023.10465461

[39] T. M. van Engers and D. M. de Vries, "Governmental transparency in the era of artificial intelligence," in *Frontiers in Artificial Intelligence and Applications*, vol. 322, 2019, pp. 33–42. [Online]. Available: https://doi.org/10.3233/FAIA190304

[40] C. Lee and K. Cha, "Fat-cat explainability and augmentation for an ai system: A case study on ai recruitment-system adoption," *International Journal of Human-Computer Studies*, vol. 171, p. 102976, 2023. [Online]. Available: https://doi.org/10.1016/j.ijhcs.2022.102976

[41] M. Vössing, N. Kühl, M. Lind, K. Främling, and G. Satzger, "Designing transparency for effective human-ai collaboration," *Information Systems Frontiers*, vol. 24, pp. 877–895, 2022.

[42] O. Adamyk, O. Chereshnyuk, B. Adamyk, and S. Rylieiev, "Trustworthy ai: A fuzzy-multiple method for evaluating ethical principles in ai regulations," in *Proceedings of the 13th International Conference on Advanced Computer Information Technologies (ACIT)*, Wrocław, Poland, 2023. [Online]. Available: https://doi.org/10.1109/ACIT58437.2023.10275505

[43] I. Barclay, H. Taylor, A. Preece, I. Taylor, D. Verma, and G. de Mel, "A framework for fostering transparency in shared artificial intelligence models by increasing visibility of contributions," *Concurrency and Computation: Practice and Experience*, vol. 33, no. e6129, 2021.

[44] M. Mylrea and N. Robinson, "Artificial intelligence (ai) trust framework and maturity model: Applying an entropy lens to improve security, privacy, and ethical ai," *Entropy*, vol. 25, no. 10, p. 1429, 2023. [Online]. Available: https://doi.org/10.3390/e25101429

[45] B. Mittelstadt, "Interpretability and transparency in artificial intelligence," in *The Oxford Handbook of Digital Ethics*, C. Véliz, Ed. Oxford University Press, 2021, online edition, accessed 17 May 2025. [Online]. Available: https://doi.org/10.1093/oxfordhb/9780198857815.013.20

[46] G. Karanxha, "Evaluating transparency in artificial intelligence systems: Adapting the z-inspection® framework for the manolo project," Bachelor's Thesis, Laurea University of Applied Sciences, Espoo, Finland, 2025.