information

Systematic Review

Illuminating Industry Evolution: Reframing Artificial Intelligence Through Transparent Machine Reasoning

Albérico Travassos Rosário [1,2,*] and Joana Carmo Dias [3,4]

[1] Instituto Politécnico de Setúbal, Escola Superior de Ciências Empresariais de Setúbal, Campus do Instituto Politécnico de Setúbal, Edifício ESTS, Estefanilha, 2914-504 Setúbal, Portugal
[2] GOVCOPP—Governance, Competitiveness and Public Policies, Campus Universitário de Santiago, Rua de S. Tiago DCSPT Room 12.3.8, 3810-193 Aveiro, Portugal
[3] Centro de Investigação em Organizações, Mercados e Gestão Industrial (COMEGI), Universidade Lusíada, 1349-001 Lisbon, Portugal; jsdias@fep.up.pt
[4] Faculdade de Economia, Universidade do Porto (FEP), R. Roberto Frias, 4200-464 Porto, Portugal
* Correspondence: alberico@ua.pt

Abstract

As intelligent systems become increasingly embedded in industrial ecosystems, the demand for transparency, reliability, and interpretability has intensified. This study investigates how explainable artificial intelligence (XAI) contributes to enhancing accountability, trust, and human–machine collaboration across industrial contexts transitioning from Industry 4.0 to Industry 5.0. To achieve this objective, a systematic bibliometric literature review (LRSB) was conducted following the PRISMA framework, analysing 98 peer-reviewed publications indexed in Scopus. This methodological approach enabled the identification of major research trends, theoretical foundations, and technical strategies that shape the development and implementation of XAI within industrial settings. The findings reveal that explainability is evolving from a purely technical requirement to a multidimensional construct integrating ethical, social, and regulatory dimensions. Techniques such as counterfactual reasoning, causal modelling, and hybrid neuro-symbolic frameworks are shown to improve interpretability and trust while aligning AI systems with human-centric and legal principles, notably those outlined in the EU AI Act. The bibliometric analysis further highlights the increasing maturity of XAI research, with strong scholarly convergence around transparency, fairness, and collaborative intelligence. By reframing artificial intelligence through the lens of transparent machine reasoning, this study contributes to both theory and practice. It advances a conceptual model linking explainability with measurable indicators of trustworthiness and accountability, and it offers a roadmap for developing responsible, human-aligned AI systems in the era of Industry 5.0. Ultimately, the study underscores that fostering explainability not only enhances functional integrity but also strengthens the ethical and societal legitimacy of AI in industrial transformation.

Keywords: artificial intelligence; explainable AI; industry; Industry 4.0; Industry 5.0; decision making

Academic Editor: Rao Mikkilineni

Received: 25 August 2025
Revised: 5 November 2025
Accepted: 7 November 2025
Published: 1 December 2025

Citation: Rosário, A.T.; Dias, J.C. Illuminating Industry Evolution: Reframing Artificial Intelligence Through Transparent Machine Reasoning. *Information* **2025**, *16*, 1044. https://doi.org/10.3390/info16121044

1. Introduction

Recent technological advancements have led to the rapid development of machine learning and artificial intelligence (AI) models capable of performing complex tasks with unprecedented precision and speed. AI innovations are transforming all sectors—from

natural language processing and image recognition to financial modelling, healthcare diagnostics, and predictive maintenance—positioning AI as a cornerstone of contemporary decision-making [1] (Felzmann et al., 2020). However, as these systems increasingly influence high-stakes decisions, their opaque "black-box" nature raises significant concerns regarding transparency, accountability, and trust. Studies such as Greene et al. (2019) [2] and Felzmann et al. (2020) [1] highlight persistent issues of bias, discrimination, and lack of interpretability in AI/ML systems, underscoring the need for mechanisms that ensure these technologies remain auditable, explainable, and ethically aligned.

This growing imperative for transparency has accelerated the emergence of Explainable Artificial Intelligence (XAI), a field focused on designing models that articulate their decision logic without compromising performance. XAI represents a paradigm shift from opaque algorithmic systems to interpretable frameworks capable of communicating reasoning processes to human users. Hollanek (2023) [3] describes this shift as "debunking the illusion, seeing through—to reveal how an object really works" (p. 2071). As defined by Das and Rad (2020) [4], XAI offers "a set of tools, techniques, and algorithms that can generate high-quality, interpretable, human-understandable explanations of AI decisions" (p. 1). Beyond technical explainability, these approaches are essential for building trust, ensuring accountability, and aligning AI systems with human cognitive and ethical expectations.

Despite increasing scholarly and industrial attention, research at the intersection of XAI and industrial transformation (Industry 4.0 and 5.0) remains fragmented and primarily technical. Most studies examine isolated applications, such as predictive maintenance, smart manufacturing, or automation, without addressing the broader theoretical, ethical, and regulatory dimensions of explainability. This fragmentation reveals a critical research gap: the lack of an integrated and systematic understanding of how XAI contributes to responsible and human-centric industrial ecosystems. Moreover, as Industry 4.0 evolves toward Industry 5.0—emphasising human-centricity, sustainability, and resilience—there is an urgent need to reconceptualise AI not merely as an autonomous decision-maker but as a collaborative and transparent partner in industrial value creation [5].

In response to this gap, the present study conducts a Systematic Bibliometric Literature Review (LRSB) guided by the PRISMA framework. This methodological approach enables a rigorous mapping of the scientific landscape to identify and synthesise research trends and dominant themes concerning XAI in industrial contexts; assess methodological and theoretical approaches that define how explainability supports accountability, fairness, and interpretability in industrial decision-making; and highlight gaps and future research directions to guide the transition from Industry 4.0 to Industry 5.0 through transparent machine reasoning [6].

By systematically integrating bibliometric and conceptual insights, this study aims to reframe artificial intelligence through transparent reasoning, positioning XAI as a key enabler of trust and ethical alignment in next-generation industrial systems.

The remainder of this paper is structured as follows: Section 2 details the materials and methods, explaining the LRSB process and PRISMA protocol; Section 3 presents the bibliometric results, including publication trends, thematic maps, and keyword networks; Section 4 discusses theoretical perspectives and emerging XAI techniques; and Section 5 concludes with the main findings, theoretical contributions, managerial implications, and directions for future research.

2. Materials and Methods

The study employs a systematic bibliometric literature review (LRSB) methodology guided by the PRISMA framework. Marzi et al. (2024) [7] describe the LRSB as a research approach that enables scholars to synthesise and explore existing knowledge paths by

spotlighting gaps and interconnections and critically assessing prior literature. As a result, this methodology facilitates a structured understanding of the field, helping to identify underexplored areas, emerging themes, and the intellectual progression of research. Combining the LRSB with the PRISMA framework enhances methodological transparency by providing clear, replicable criteria for literature selection, screening, and inclusion (Supplementary Materials) [8,9]. The integrated methodology helps explore how researchers have conceptualised, debated, and integrated XAI across scholarly discourse.

Rosário and Dias [10–12] describe the LRSB technique as offering a more structured and thorough approach to exploring a research field than what is typically found in traditional literature reviews. Instead of gathering a wide range of studies, this method focuses on carefully selecting works that directly address the core research question, all while maintaining a high level of transparency. This thoughtful selection process allows for a deep analysis of each study's methods, the strength of its conclusions, and its overall contribution to the field.

The LRSB technique follows a clear and systematic framework, guiding researchers step by step in filtering and evaluating sources to ensure both their credibility and relevance. According to the authors, this process is divided into three main phases, encompassing six distinct steps, as outlined in Table 1.

Table 1. Process of systematic LRSB.

Fase	Step	Description
Exploration	Step 1	formulating the research problem
	Step 2	searching for appropriate literature
	Step 3	critical appraisal of the selected studies
	Step 4	data synthesis from individual sources
Interpretation	Step 5	reporting findings and recommendations
Communication	Step 6	Presentation of the LRSB report

Source: own elaboration.

This study used the Scopus database to identify and select relevant academic sources, taking advantage of its strong reputation within the scholarly community. The exclusive focus on Scopus was a deliberate choice, based on its wide subject coverage, strict indexing criteria, and powerful analytical tools. As one of the most comprehensive platforms for peer-reviewed research, Scopus brings together content from respected journals, books, and conference proceedings across a broad range of disciplines. Its features for citation tracking, bibliometric analysis, and keyword mapping allow for an in-depth exploration of research trends and scholarly impact—critical elements for developing a robust and methodologically sound literature review. In addition, Scopus's standardised metadata and consistent indexing practices help ensure transparency and make the research process easier to replicate. Given these advantages, restricting the review to Scopus sources was a strategic decision aimed at maintaining analytical rigour and reliability in this academic investigation.

However, relying exclusively on Scopus is also a limitation, as it risks omitting relevant studies, particularly regional contributions, non-English publications, or very recent works not yet indexed. While this constraint was acknowledged, future research could expand the scope by incorporating complementary databases such as Web of Science, Google Scholar, or IEEE Xplore to ensure broader coverage and mitigate potential selection bias.

The inclusion and exclusion criteria were designed to ensure methodological transparency and thematic coherence. Only peer-reviewed academic publications, including

journal articles, conference papers, book chapters, and books, were considered. Preprints, theses, and grey literature were excluded to maintain academic quality. Duplicate records were systematically removed through metadata comparison and manual verification. Although the study excluded non-peer-reviewed materials, this may have constrained the representation of emerging discussions still in progress.

While the step-by-step process ensures procedural clarity, the authors acknowledge that the methodological choices, particularly the database restriction and exclusion of non-peer-reviewed sources, may influence the scope and balance of the review, potentially favouring English-language and mainstream academic outputs. Nevertheless, this approach enhances replicability and strengthens alignment with the study's objectives by focusing on high-impact, peer-reviewed contributions.

The literature search process began with identifying the database, in this case, the Scopus database. This selection was based on Baas et al. (2020) [13], who recognise Scopus as one of the "largest curated abstract and citation databases," offering a wide range of high-quality resources from scientific journals, books, and conference proceedings. The initial search began with the keyword "artificial intelligence," resulting in 689,969 documents. Adding the keyword "explainable AI" reduced these search results to 5496. The researcher then added the keyword "industry," which further reduced the document results to 467. Keywords "industry 4.0" further reduced the results to 45, which then increased to 49 after adding "industry 5.0." Finally, the researcher limited the search to the exact keyword "decision making," resulting in 98 sources that were synthesised in the final reporting. The Boolean query used was (TITLE-ABS-KEY ("artificial intelligence") AND TITLE-ABS-KEY ("explainable AI") AND TITLE-ABS-KEY ("Industry")) AND (LIMIT-TO (EXACTKEYWORD, "Industry 4.0") OR LIMIT-TO (EXACTKEYWORD, "Industry 5.0") OR LIMIT-TO (EXACTKEYWORD, "Decision Making")).

To ensure that the final selection of studies was both relevant and methodologically solid, the research applied clearly defined inclusion and exclusion criteria (see Table 2). The review focused exclusively on peer-reviewed publications that examined Explainable AI in industrial contexts, particularly within the evolution from Industry 4.0 to Industry 5.0. Any works that did not address this intersection were excluded to maintain thematic focus and methodological consistency.

Table 2. Screening Methodology.

Database Scopus	Screening	Publications
Meta-search	Keyword: artificial intelligence	689,969
First Inclusion Criterion	Keyword: artificial intelligence, explainable AI	5496
Inclusion Criteria	Keyword: artificial intelligence, explainable AI, Industry	467
	Keyword: artificial intelligence, explainable AI, Industry, Industry 4.0	45
	Keyword: artificial intelligence, explainable AI, Industry, Industry 4.0, Industry 5.0	49
	Keyword: artificial intelligence, explainable AI, Industry, Industry 4.0, Industry 5.0, Decision Making	98
Screening	Keyword: artificial intelligence, explainable AI, Industry, Industry 4.0, Industry 5.0, Decision Making Until June 2025	98

Source: own elaboration.

This careful filtering process helped ensure that the selected literature was academically rigorous and directly aligned with the study's objectives, namely, to map research trends, identify dominant themes, and reveal knowledge gaps concerning the role of XAI

in industrial transformation. In this way, the methodological design directly supports the research goal of clarifying how explainability contributes to accountability, interpretability, and trust in industrial decision-making systems.

A thorough analysis of the selected materials was conducted using a structured approach inspired by Rosário and Dias [10–12], with careful attention given to both the content and underlying themes. Each study was evaluated for its methodological quality, theoretical contribution, and relevance to the intersection of XAI and industrial transformation. While this approach strengthens analytical depth, it also highlights the potential for bias introduced by database and language constraints—an issue that future studies should address by adopting multi-source strategies and multilingual screening protocols.

A visual representation of this selection process is provided in Figure 1.

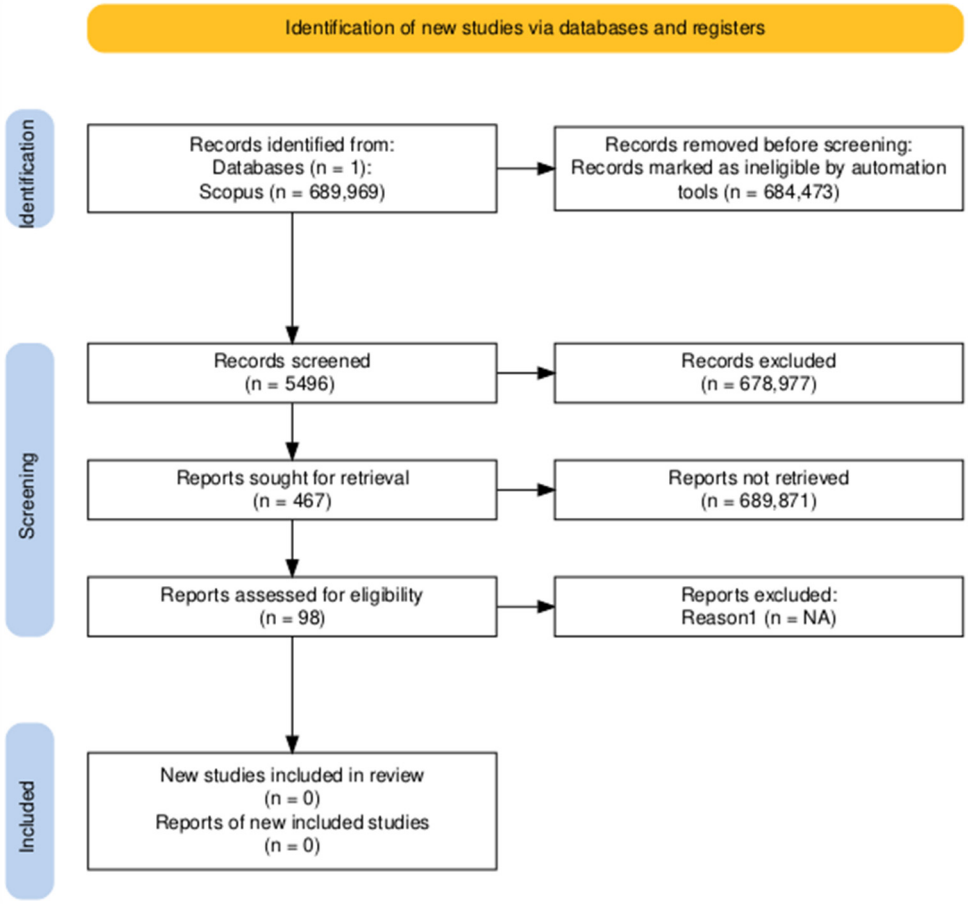

Figure 1. PRISMA 2020 Flow Diagram of Study Selection Process [9].

Each source was critically evaluated for its relevance, methodological robustness, and publication quality. In total, 98 academic and scientific documents were analysed using a combination of narrative synthesis and bibliometric techniques, following the methodological guidelines of Rosário and Dias [10–12]. This dual approach allowed for a comprehensive exploration of patterns, relationships, and research trends that directly address the study's objectives.

3. Publication Distribution

This section presents the bibliometric distribution and analytical trends derived from the 98 documents included in the systematic bibliometric literature review. The purpose of this section is to provide an interpretive overview of how research on Explainable Artificial Intelligence (XAI) has evolved within the context of industrial transformation, and how

this distribution relates to the broader objectives of the study—namely, to understand how explainability contributes to accountability, transparency, and human-centric decision-making across Industry 4.0 and Industry 5.0.

3.1. Search Parameters and Dataset Scope

The bibliometric dataset was compiled exclusively from the Scopus database, chosen for its robust citation tracking, structured metadata, and wide disciplinary coverage. The search was performed in June 2025, using the following Boolean query:

TITLE-ABS-KEY ("artificial intelligence") AND TITLE-ABS-KEY ("explainable AI") AND TITLE-ABS-KEY ("industry" OR "industry 4.0" OR "industry 5.0") AND TITLE-ABS-KEY ("decision making")

To ensure transparency and replicability, the following parameters were applied:

- Source type: Peer-reviewed journal articles, books, book chapters, and conference papers;
- Language: English;
- Publication years: ≤June 2025;
- Subject areas: Computer Science, Engineering, Decision Sciences, Business, Management & Accounting;
- Inclusion criteria: Documents addressing explainability in industrial or decision-making contexts;
- Exclusion criteria: Preprints, theses, and grey literature were excluded to maintain quality; duplicates were removed through metadata comparison and manual verification.

This deliberate configuration ensured methodological rigour but also introduced inherent limitations, notably the exclusion of regional or non-English contributions that might provide complementary insights. The implications of this bias are discussed in subsequent subsections.

3.2. Temporal Evolution of Publications

Figure 2 illustrates the annual growth of publications between 2018 and June 2025. The pattern reveals a pronounced increase beginning in 2020, culminating in a publication peak in 2024 (44 papers), before a moderate decline in 2025. This temporal trajectory mirrors the intensification of scholarly and policy interest in ethical and transparent AI, particularly following the consolidation of frameworks such as the European Union Artificial Intelligence Act (EU AI Act).

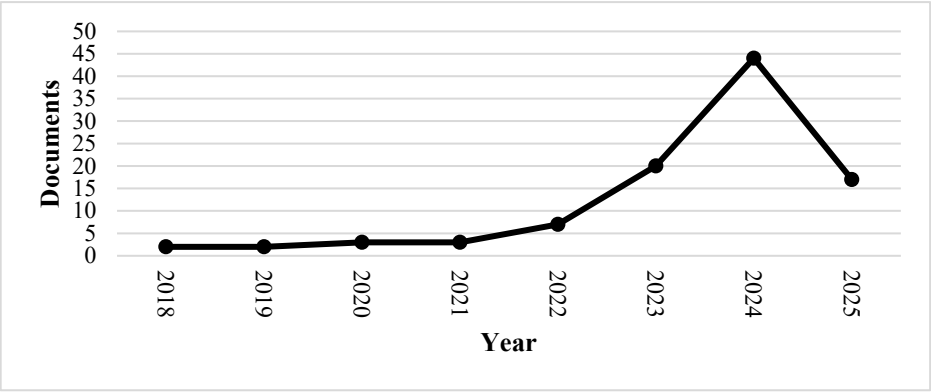

Figure 2. Documents by year.

Rather than indicating a mere quantitative expansion, this evolution reflects a conceptual transformation in the literature—from a focus on algorithmic optimisation toward human-centred, interpretable, and ethically responsible AI. The rise in 2024 corresponds to the convergence of discussions around trustworthiness, regulation, and Industry 5.0's

human-centric paradigm, signalling a transition in research priorities from automation to collaboration.

3.3. Geographic Distribution and Research Biases

Figure 3 and Table 3 present the global distribution of publications by country. The dataset shows a clear concentration of research activity in technologically advanced economies, with India (66), the United States (44), Italy (32), Greece (31), and Germany (28) leading in publication output. This pattern suggests that XAI research is strongly linked to industrial digitalisation capacity, national AI strategies, and investment in innovation ecosystems.

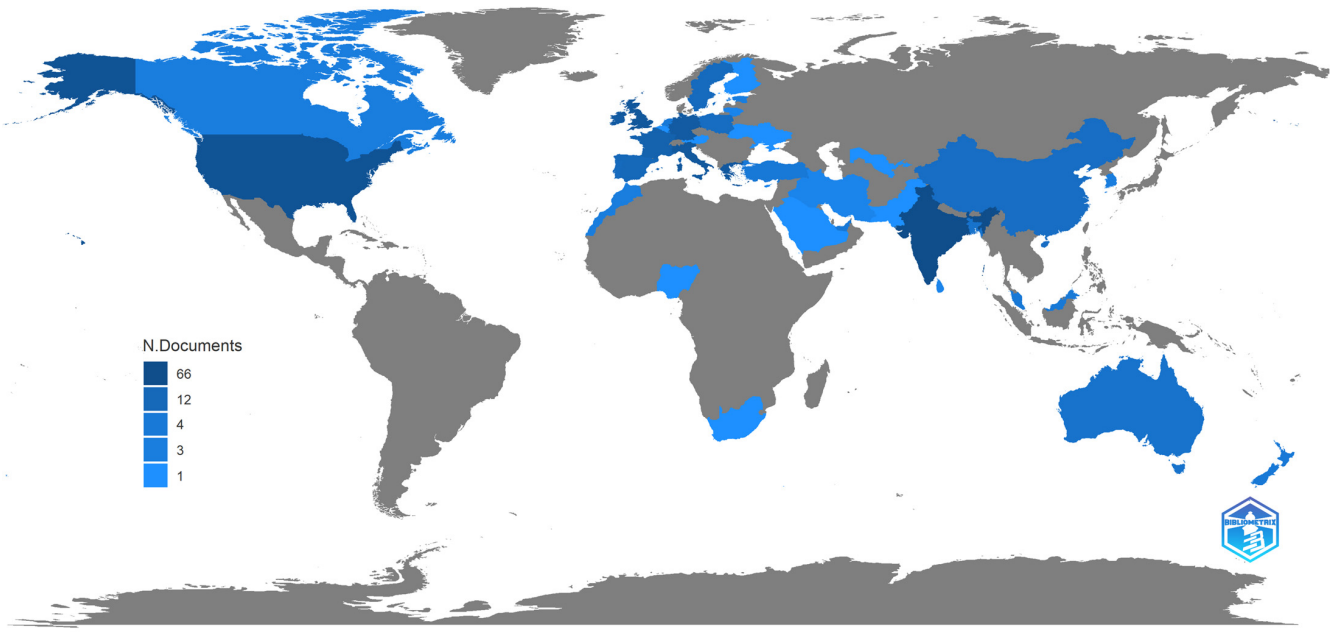

Figure 3. Scientific production by country.

Table 3. Top 10 countries by number of publications.

Country	Number of Publications
India	66
USA	44
UK	33
Italy	32
Greece	31
Germany	28
Ireland	26
France	17
Spain	12
Portugal	10

Source: own elaboration.

However, this distribution also reveals regional asymmetries that align with the known Scopus indexing bias, which tends to privilege English-language journals and Western publication outlets. Consequently, research produced in Latin America, Africa, and parts of Asia may be underrepresented, limiting the inclusiveness of the global perspective. Despite these imbalances, the growing participation of European and Asian countries indicates that explainable AI has become a globalised yet unevenly distributed research domain.

These findings reveal how the geopolitics of innovation influences the production of knowledge in XAI. Nations with robust AI policies and industrial transformation agendas—such as India, the USA, and EU member states—serve as intellectual hubs driving the global conversation on ethical AI deployment. This geographical concentration also reflects how industrial maturity correlates with research visibility, an insight that connects directly to this review's focus on the evolving industrial paradigm.

3.4. Core Journals and Thematic Concentration

Applying Bradford's Law, the analysis identified a core cluster of journals that account for approximately 10% of total publications, indicating a growing consolidation of XAI research around key scholarly outlets (Figure 4). The most prominent among these are IEEE Access, Procedia Computer Science, Ceur Workshop Proceedings, and Applied Sciences.

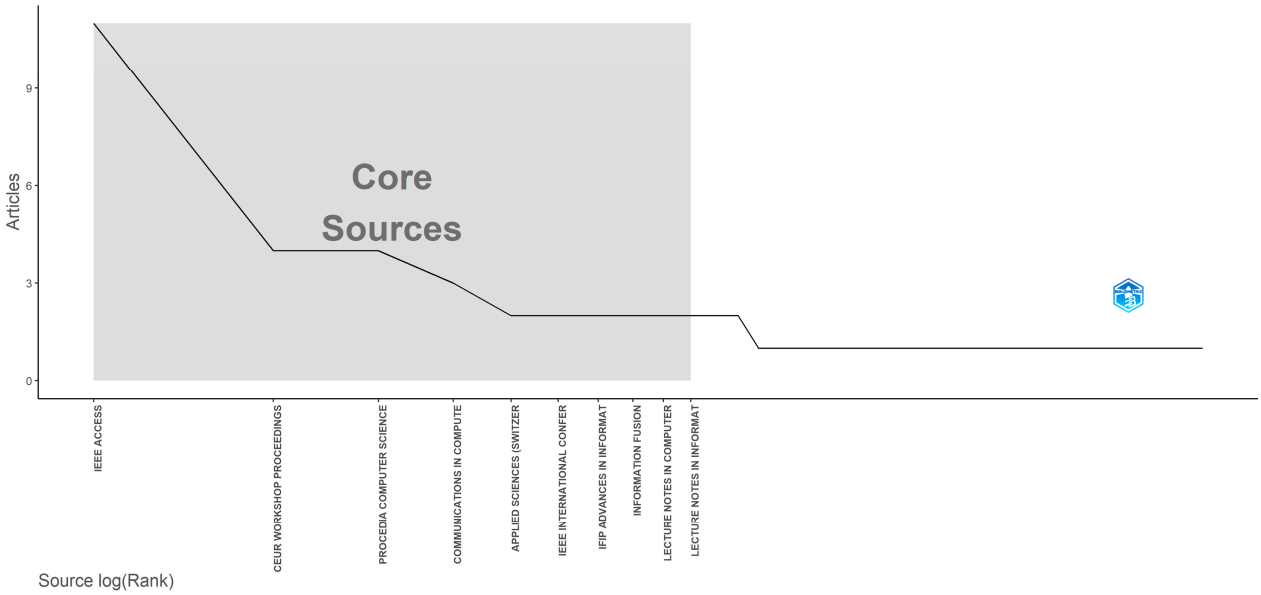

Figure 4. Core sources by Bradford's law (2015–2025).

This concentration suggests a maturing research domain in which publication venues have become specialised platforms for interdisciplinary exchange between AI engineering, ethics, and industrial management. The consolidation of these journals underscores their role in setting the academic agenda for explainability in industrial ecosystems—particularly in relation to predictive maintenance, trust in automation, and hybrid human–machine collaboration.

The stabilisation of these outlets highlights the emergence of a cohesive scholarly community, aligning with one of the review's central objectives: to trace the intellectual architecture and publication dynamics that underpin XAI's integration into Industry 4.0 and 5.0 frameworks.

3.5. Disciplinary Landscape and Research Impact

The 98 documents analysed spanned 18 Scopus subject categories, confirming the field's interdisciplinary nature. Most publications are concentrated in Computer Science (79) and Engineering (46), followed by Mathematics (17), Decision Sciences (14), and Business, Management, and Accounting (9). This distribution demonstrates that explainability is no longer an exclusively technical challenge but an emerging managerial and socio-technical issue central to decision-making in industrial environments.

The most cited article, "A Review of Trustworthy and Explainable Artificial Intelligence (XAI)" (IEEE Access), has received 114 citations, exemplifying the field's emphasis on integrating trust, transparency, and ethical alignment into AI development. As of June 2025, the dataset accumulated 1068 citations (Appendix A), with an h-index of 18, confirming a high degree of intellectual influence and topic maturity.

Figure 5 illustrates the citation trajectory, showing steady growth since 2018. This increase is not merely numerical; it signifies the progressive diffusion of explainability principles into industrial AI applications and the consolidation of XAI as a fundamental dimension of technological governance.

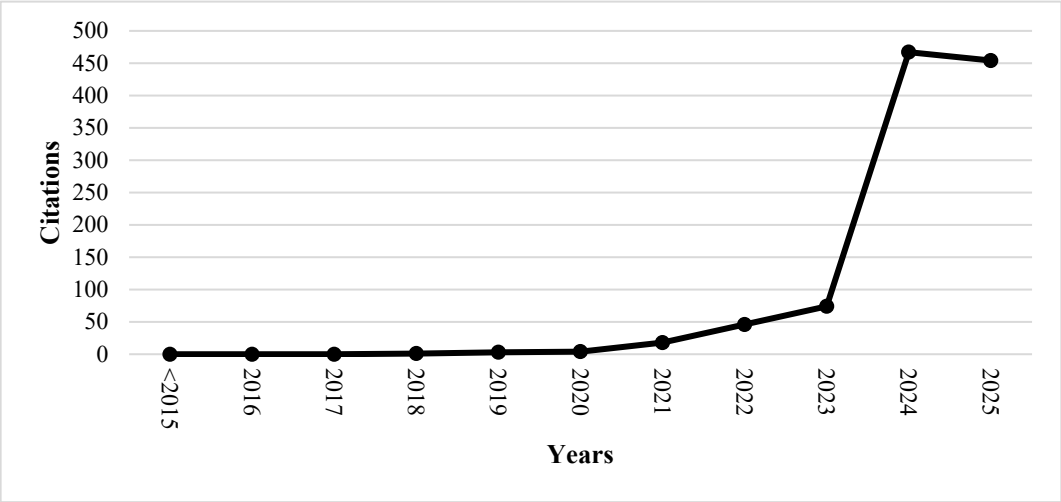

Figure 5. Evolution of citations between ≤2015 and 2025.

3.6. Keyword Co-Occurrence and Thematic Networks

To examine the intellectual structure of the field, a keyword co-occurrence analysis was conducted using VOSviewer (version 1.6.18). Core keywords included artificial intelligence, explainable AI, industry 4.0, industry 5.0, and decision-making.

Figure 6 visualises the co-occurrence network, where node size represents keyword frequency and proximity indicates thematic relatedness. The clustering reveals three main conceptual domains:

- Technical explainability (machine learning, deep learning, neural networks);
- Industrial application (predictive maintenance, resource optimisation);
- Ethical and organisational dimensions (trust, transparency, accountability).

The convergence of these domains confirms that explainability operates as a bridging construct, connecting algorithmic logic with human interpretability—precisely the relationship explored in this review.

3.7. Three-Field Plot: Authors, Keywords, and References

To map the intellectual flow of knowledge, a three-field plot (Figure 7) was generated using the Bibliometrix package (version 5.1.1). The plot links authors (AU), cited references (CR), and keywords (DE), highlighting influential contributors and conceptual overlaps.

The analysis shows that the most prolific authors are associated with recurring keywords such as explainable AI, Industry 4.0, trustworthy systems, and decision support, indicating a strong convergence between research productivity and thematic focus.

This mapping reveals that the field's intellectual structure is dominated by cross-disciplinary collaborations between computer scientists and industrial engineers, reinforcing the importance of XAI as a transversal competence in the transition to Industry 5.0.

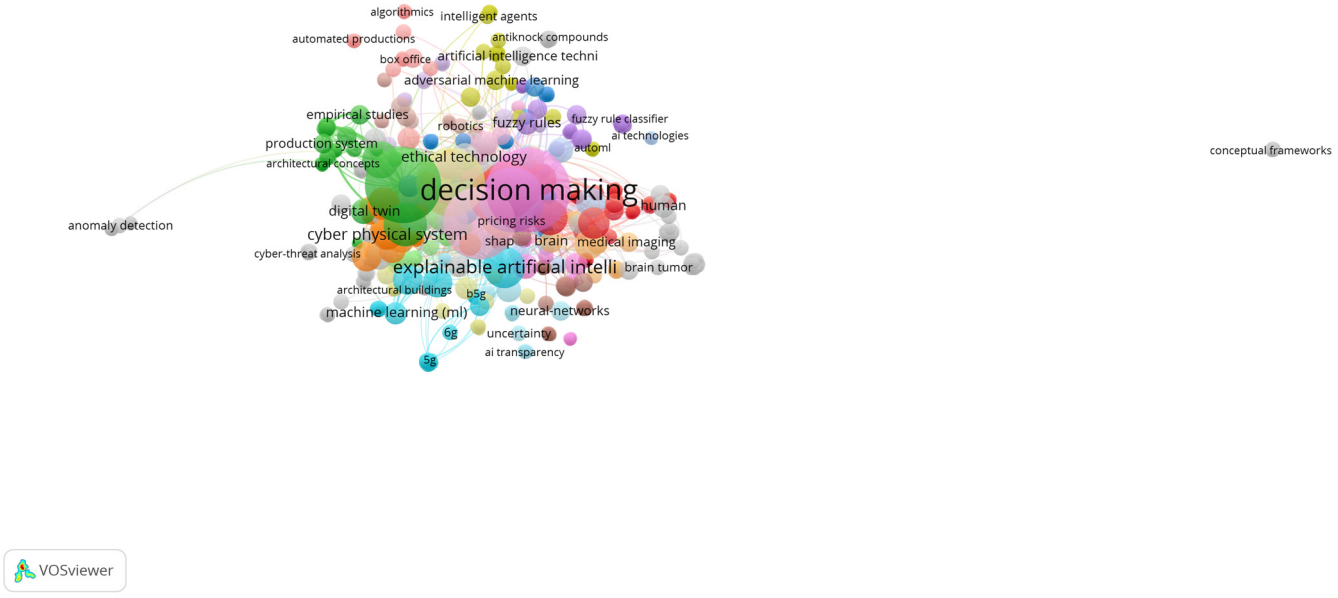

Figure 6. Network of all keywords.

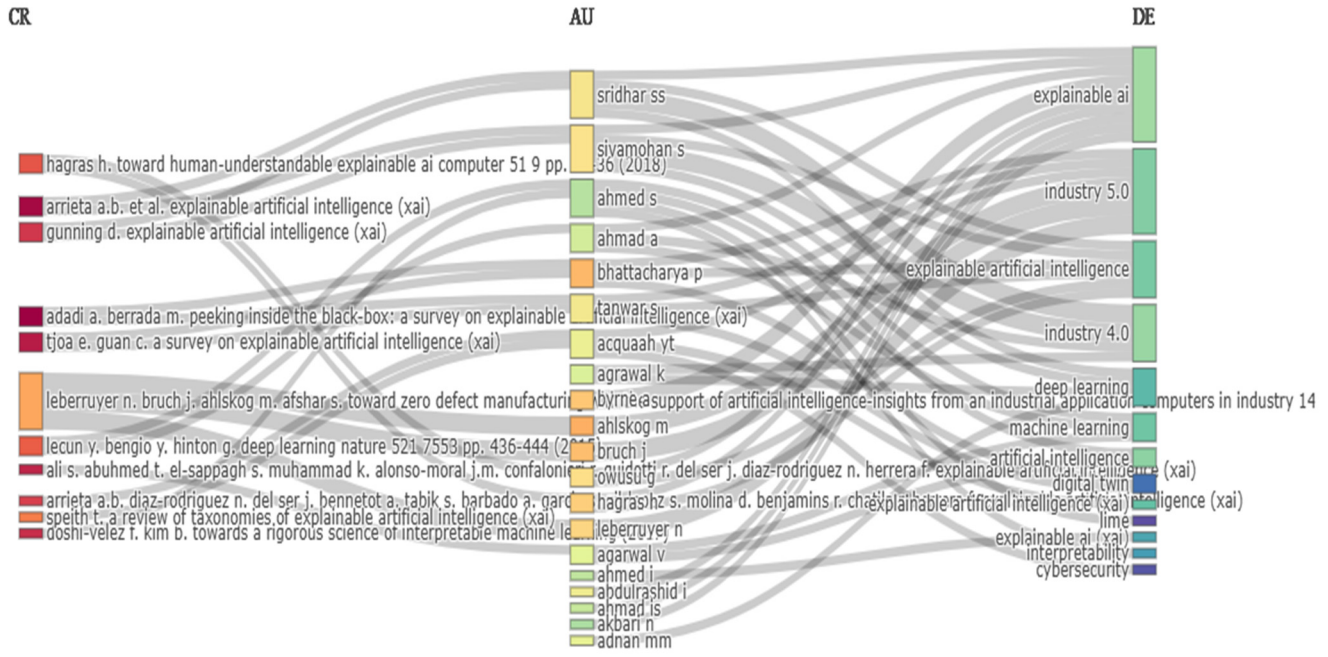

Figure 7. Three fields plot analysis (AU = authors, CR = references, DE = author keywords).

3.8. Thematic Map Analysis

A thematic map (Figure 8) was produced to examine the maturity and centrality of research topics. The map is divided into four quadrants—motor themes, basic themes, emerging/declining themes, and niche themes—based on the dimensions of density and centrality.

Motor themes (upper-right quadrant): "Industry 4.0," "deep learning," "predictive maintenance," and "artificial intelligence learning"—conceptually mature and core to the field.

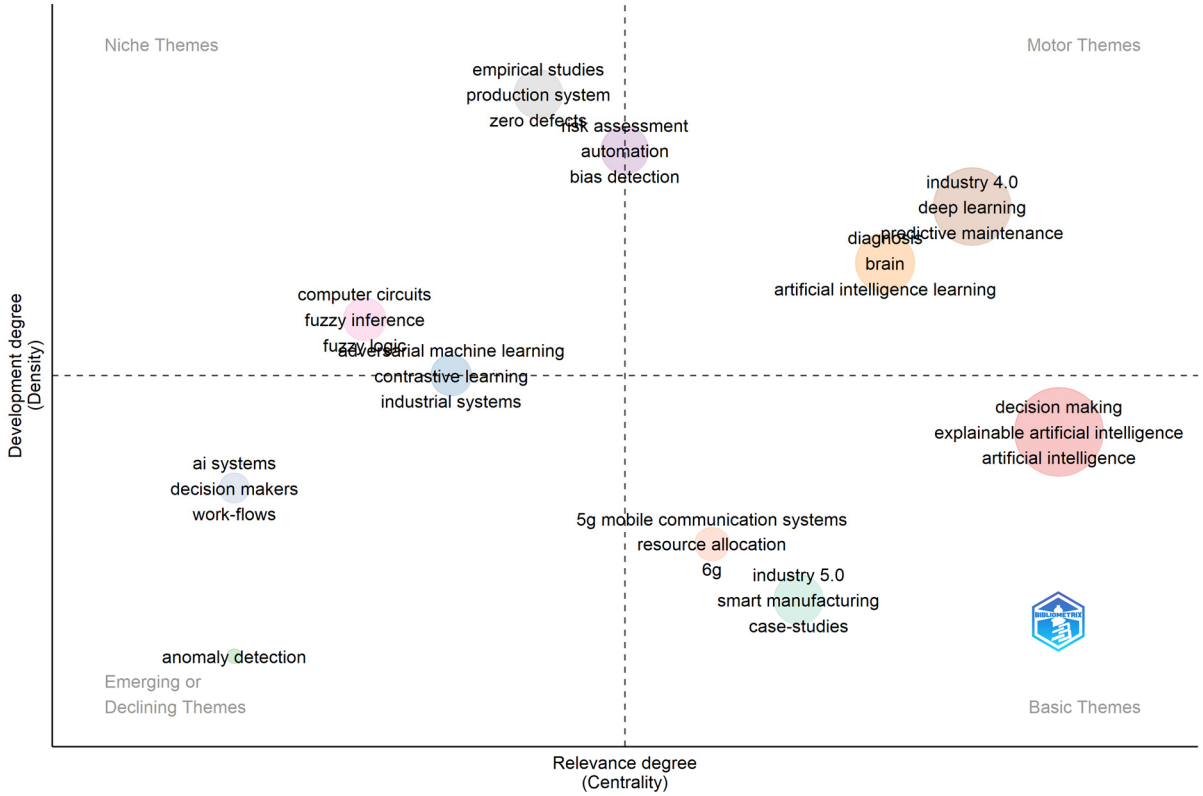

Figure 8. Thematic map analysis.

Basic and transversal themes (lower-right quadrant): "Industry 5.0," "smart manufacturing," "resource allocation," and "case studies," which represent foundational areas for future research expansion.

Niche themes (upper-left quadrant): "automation," "bias detection," and "zero defects," showing strong internal coherence but limited broader influence.

Emerging/declining themes (lower-left quadrant): "decision makers," "workflows," and "anomaly detection," suggesting potential new directions or fading relevance.

This structure demonstrates the dynamic evolution of XAI research from purely computational themes to socio-technical and managerial applications, reinforcing the interdisciplinary character central to this review.

3.9. Co-Citation Network Analysis

Finally, a co-citation network (Figure 9) was developed to visualise how foundational works are interlinked through shared citation patterns. The central cluster is formed by seminal studies on trustworthy AI, explainability techniques, and ethical governance, which constitute the intellectual backbone of the field.

Peripheral nodes represent emerging research streams, such as fairness in AI, hybrid human–machine collaboration, and sustainability in Industry 5.0. The presence of these clusters illustrates that while XAI research is theoretically consolidated, it remains open to diversification and innovation

The co-citation structure captures a research community that is methodologically robust yet conceptually expanding—a field moving toward synthesis between technological precision and human accountability, which is the defining argument of this paper.

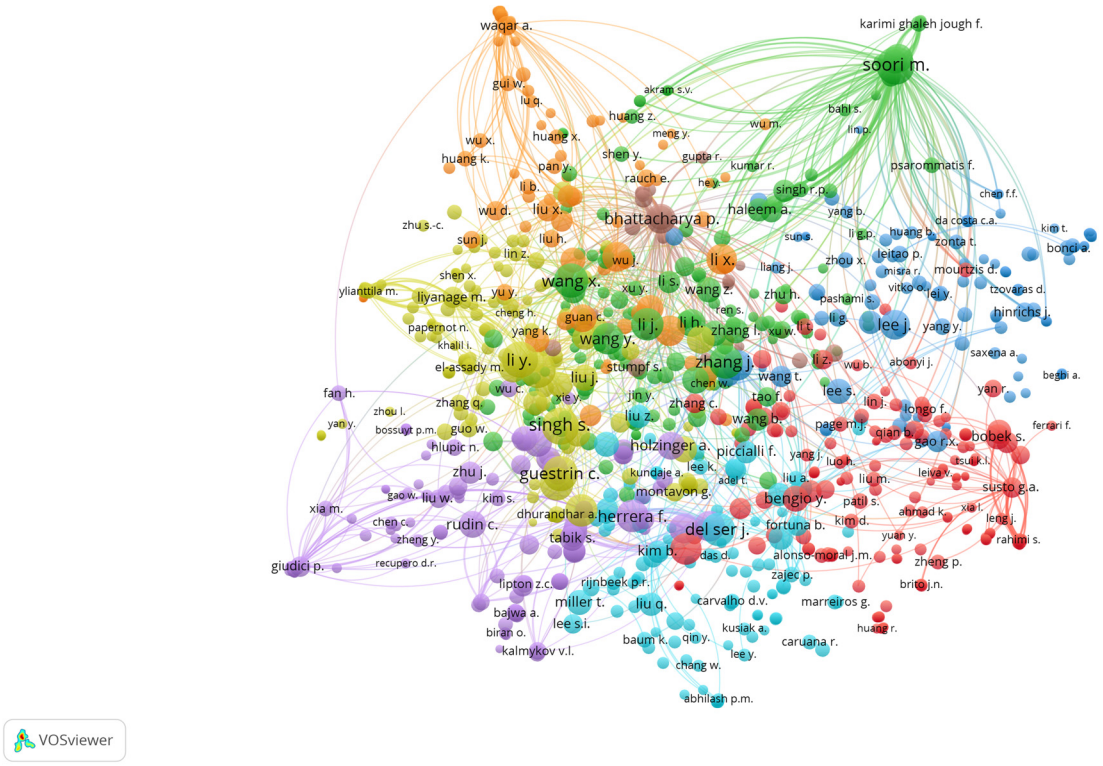

Figure 9. Network of co-citation.

4. Theoretical Perspectives

Opacity is a major problem in AI/ML decision-making systems. Facchini and Termine (2021) [13] explain that these opaque models make it difficult for users to understand how they work, interpret decisions at various levels, and assess their behaviours against ethical and scientific norms. The inability to trace how an algorithm arrived at a particular outcome can undermine accountability, hinder error detection, and exacerbate social biases embedded in training data [14,15] (Chamola et al., 2023; Ganguly & Singh, 2023). As a result, Facchini and Termine (2021) [13] note that many philosophers and social scientists have begun researching and interpreting the opacity issue and its implications, while engineers are working to create explainable AI systems that mitigate the opacity. These XAI models aim to improve transparency and make AI/ML systems more understandable to humans.

4.1. Goals for Pursuing Explainability in XAI

AI systems have become increasingly embedded in decision-making processes. As a result, the demand for explainability has expanded beyond technical performance to encompass ethical, social, and legal considerations. In this case, Miltiadou et al. (2023) [16] explain that XAI is not just about understanding how an algorithm works but also about designing systems that align with human values, societal expectations, and institutional accountability. The goals outlined below represent the multi-dimensional motivations driving the pursuit of explainability in AI development and deployment:

4.1.1. Trustworthiness

Trust is foundational for AI adoption in decision-critical settings [17]. Explainability contributes by showing how and why outputs are produced, enabling users to justify reliance on the system [18,19]. Still, trust is contingent: it depends on whether explanations are understandable to the intended audience, anchored in domain norms, and demonstrably useful for task performance [20]. When models are opaque or appear biased or data-

dependent without justification, measurable trust indicators deteriorate (e.g., adoption rates, confidence surveys, efficiency metrics) [21,22]. In short, explainability is necessary but insufficient for trust; it must be paired with credibility, usability, and contextual fit.

4.1.2. Intelligibility

Intelligibility concerns the user's ability to grasp what a model does and why it behaves as it does, without needing to inspect algorithmic internals [23–25]. It has been linked to improvements in accuracy, decision latency, learning curves, and error reduction, though these gains vary with user training, domain knowledge, and interface quality [26,27]. Effective intelligibility helps users build stable, verifiable mental models rather than mere procedural familiarity, supporting collaboration and handoffs between human and machine.

4.1.3. Transparency

Transparency is often positioned as a foundational mechanism for explainability, offering visibility into the internal structure, logic, and processes of AI systems. Arora and Gajjar (2025) [28] define transparent systems as those that openly disclose their architecture, parameters, data sources, and decision logic, thereby enabling external auditing, interrogation, and critical evaluation. While this framing aligns transparency with accountability, debugging, and regulatory compliance, its practical implementation raises deeper tensions between interpretability, complexity, and usability.

Not all transparency is equally useful or usable. Arrieta et al. (2020) [23] propose three conceptual levels—simulatability, decomposability, and algorithmic transparency—as pathways for making models understandable. Yet these levels differ significantly in their cognitive demands and their implications for different stakeholders. What counts as transparent for a developer may remain opaque to a regulator, operator, or end-user.

Simulatability

Simulatability focuses on whether a person can mentally simulate or predict a model's behaviour based on its inputs, an ability often constrained by the model's complexity. Arrieta et al. (2020, p. 10) [23] define it as the human capacity to "think through" the model, making simplicity a prerequisite. Models like decision trees or linear regressions are typically cited as fulfilling this condition [29,30] (Beshaw et al., 2025; Alexander et al., 2024), but this introduces a trade-off: increasing model complexity to improve accuracy often reduces simulatability.

While simulatability can be measured through user prediction accuracy or time-to-decision metrics, these indicators oversimplify the problem. Users may accurately guess outputs without grasping why a model made a decision, raising concerns about false interpretability. Furthermore, simulatability benefits vary depending on task complexity and user expertise, limiting its generalizability as a universal design goal.

Decomposability

Decomposability refers to the interpretability of individual model components, such as inputs, parameters, and internal computations. Arrieta et al. (2020) [23] argue that decomposable models allow users to assess the contribution of each part to the overall output. Logistic regression, for example, facilitates understanding through direct interpretation of feature weights [31,32] (Chander et al., 2018; Bajpai et al., 2025).

Yet decomposability also faces critical limitations. It assumes that users can meaningfully isolate and understand these components, which may not hold in complex, high-dimensional models. While decomposability may improve debugging and allow traceable adjustments, the cognitive load involved can restrict its practical utility. In regulated environments, metrics like intervention rates or error-diagnosis speed provide quantifiable

benefits, but they do not resolve the fundamental issue: transparency at the component level does not guarantee comprehension at the systemic level.

Algorithmic Transparency

Algorithmic transparency concerns the visibility of the model's training and optimisation processes—including objective functions, convergence criteria, and regularisation methods. Arrieta et al. (2020) [23] suggest that this level of transparency is essential for assessing reproducibility, stability, and fairness [33,34] (Fares et al., 2023; Roy et al., 2023). Yet disclosing such technical details often favours expert users while excluding those without formal training.

Measurable indicators, such as reproducibility audits, documentation scores, and variance analyses, may signal algorithmic transparency, but they tell us little about how stakeholders interpret or act upon that information. In many cases, users are asked to place trust in documentation they cannot independently verify. Moreover, when transparency in training processes is insufficient, even highly interpretable models may be rejected as unreliable or misaligned with institutional expectations.

Taken together, the promise of transparency in explainable AI rests on its ability to make models accountable and interpretable, but its effectiveness is shaped by who is interpreting, for what purpose, and under what constraints. Simulatability prioritises cognitive accessibility but may sacrifice predictive power; decomposability enables component-level insight but struggles with systemic complexity; algorithmic transparency offers procedural visibility but often excludes non-technical users. Recognising these tensions is essential for designing transparency strategies that are not only technically sound but also contextually meaningful and socially actionable.

4.1.4. Comprehensibility

Comprehensibility focuses on cognitive alignment: explanations should resonate with users' reasoning styles, vocabulary, and domain frames [35,36]. Clinicians, for instance, often prefer clinical rationales over abstract statistics [37,38]. Techniques such as natural-language explanations, guided visuals, and step-through interfaces can improve accessibility [39,40]. The design challenge is balance: over-simplification can induce false confidence, while excessive technicality can overload or exclude users.

4.1.5. Causality

Causal explanations move beyond association to clarify mechanisms and "what-if" scenarios, enabling counterfactual and intervention-oriented reasoning [41,42]. This can increase actionability and perceived trustworthiness when interventions are required [43]. However, valid causal accounts depend on strong assumptions about data, identifiability, and feasibility; if left implicit, they may mislead. Causal XAI therefore demands explicit statement of assumptions and empirical checks.

4.1.6. Transferability

Transferability asks whether explanations retain their meaning across datasets, contexts, and user groups [23]. Consistent explanatory patterns help standardise operations and support governance across units or jurisdictions [44–46]. Yet over-generalised explanations may lose local nuance, while highly tailored ones can be hard to reuse. Designing for transferability means balancing stability with contextual adaptation and routinely validating whether explanations still "travel well."

4.1.7. Informativeness

Informativeness refers to explanatory depth and practical utility: which features mattered, how similar cases were treated, what uncertainty surrounds the output [47,48]. More detail is not always better; too much technical content can obscure signal, while oversimplification hides trade-offs and limitations [49,50]. Targeting the "right" level requires profiling tasks and audiences, then calibrating explanations to decision stakes and time constraints.

4.1.8. Interactivity

Interactivity reframes explanation as dialogue: users query, test, and manipulate inputs while the system responds with tailored feedback [23,51]. Interfaces, dashboards, and conversational agents can improve understanding, engagement, and retention [36,52]. Poorly designed interaction, however, may produce confusion, over-fitting to user probes, or hidden drift in explanation behaviour, which raises issues of traceability, versioning, and reproducibility [53,54]. Interactive XAI should therefore include guidance, guardrails, and logs.

4.1.9. Improve Model Bias Understanding and Fairness

Explainability can surface disparate impacts and proxy effects, revealing how features drive outcomes across groups [55,56]. Yet visibility alone does not remediate bias. Fairness claims are plural and sometimes conflicting (distributive vs. procedural criteria) [57] (Figure 10); progress requires organisational mechanisms that can act on evidence (e.g., monitoring, escalation paths, corrective retraining) [58]. Explainability supports fairness when embedded in governance, not as a stand-alone metric.

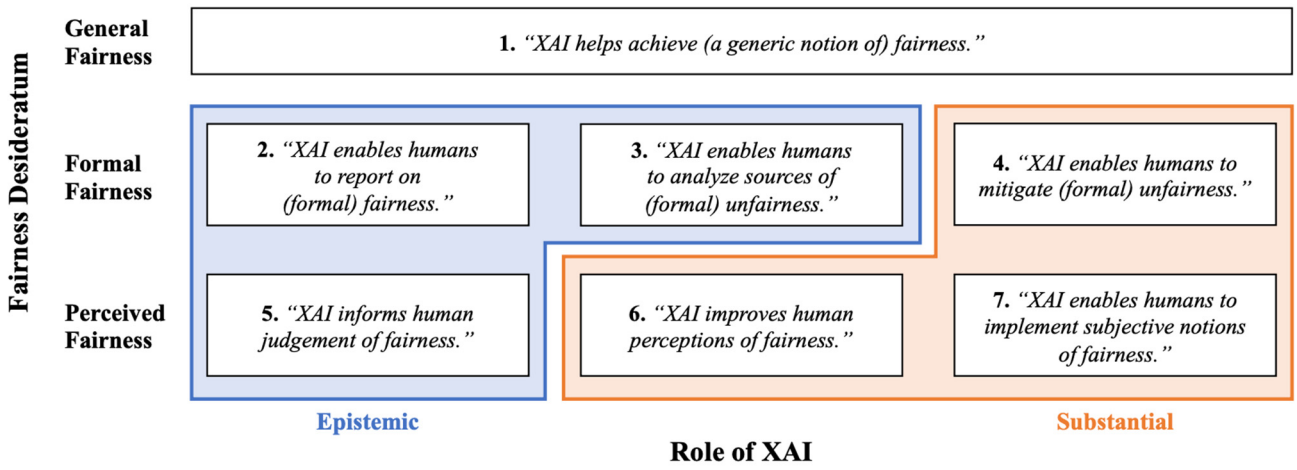

Figure 10. Seven claims of fairness related to XAI models [57] (Deck et al., 2023).

4.1.10. Accessibility

Accessibility extends explainability across different literacies, languages, and cognitive profiles [23,59]. Without inclusive design, explanatory tools can reproduce inequalities or assume unrealistic baseline competencies [60,61]. Practical measures include multilingual content, layered explanations, and participatory testing with diverse users to ensure equitable comprehension and use.

4.1.11. Privacy Awareness

Explanations can inadvertently expose sensitive attributes, proprietary logic, or attack surfaces [62]. Privacy-aware XAI therefore balances disclosure with protection via abstrac-

tion, access control, and privacy-preserving techniques, coupled with institutional oversight and clear protocols [63,64]. The goal is to remain informative enough for accountability while sufficiently bounded for confidentiality and safety.

4.1.12. Regulatory Compliance

With emerging frameworks (e.g., EU AI Act), explainability has become a legal obligation tied to traceability, auditability, and non-discrimination [65,66]. Documentation, audit logs, and validation artefacts help operationalise compliance [67,68], but risk drifting into box-ticking if they prioritise auditor needs over user understanding. Mature practice integrates compliance with human-centred explanation so systems are both verifiably accountable and genuinely interpretable.

4.2. XAI Techniques for Improving Explainability

Researchers and developers have designed a wide range of technical strategies to achieve the goals of explainability, such as trust, transparency, comprehensibility, and fairness. These techniques aim to make complex machine learning (ML) and AI models more understandable to human users by providing insight into how predictions are made, what factors influence decision outcomes, and how models behave under various scenarios [69,70] (Puthanveettil Madathil et al., 2025; Kotriwala et al., 2021). However, their effectiveness varies across industrial settings. In process industries, such as energy, chemicals, or pharmaceuticals, explainability focuses on continuous variable monitoring and causal relationships, while in discrete manufacturing, such as automotive or electronics, it centres on real-time traceability, quality control, and decision support. The following subsections critically evaluate major XAI techniques with respect to their industrial applicability, human-centric outcomes, and regulatory alignment:

4.2.1. Counterfactual Reasoning

Counterfactual reasoning shows how outcomes would differ if specific input features changed [71]. For example, a counterfactual explanation might state, *"If the requested loan amount were lower, approval would have been granted"* [71]. This aligns well with human reasoning because it illuminates cause-and-effect relationships. Effective counterfactuals require identifying minimal and feasible feature adjustments [72] and are widely applied in predictive maintenance and process optimisation, helping engineers simulate adjustments in operating conditions. In discrete manufacturing, they assist quality and inspection tasks by highlighting how small configuration shifts influence defect likelihood.

Prolixity

Prolixity refers to the tendency of counterfactual explanations to include too many feature changes, resulting in unnecessarily long or overly complex explanations. When a counterfactual modifies numerous input variables to achieve a different output, it can overwhelm users with information, making the explanation harder to follow and less actionable [72] (Keane et al., 2021). For example, if a loan application is denied and the system suggests that ten different financial attributes need to change to reverse the decision, the user may struggle to discern which features are most critical or realistic to adjust. Prolix explanations compromise the core purpose of counterfactual reasoning, clarity and simplicity, by failing to prioritise minimal changes that have maximum impact [73]. In practice, addressing prolixity involves optimising for sparse and focused counterfactuals that highlight the most relevant variables, improving both the intelligibility and usefulness of the explanation.

Sparcity

This problem is associated with the argument that good explainable counterfactuals need to be sparse. Keane and Smyth [73] (2020) indicate that this perspective is often linked to the limitations of human working memory. For example, the scholars argue that learners in concept learning prefer single-feature changes over multi-feature changes because it makes learning easier. This means counterfactuals are more likely to be effective when they focus on the most salient change necessary to alter a model's decision. Sparse counterfactuals reduce cognitive load and support clearer human reasoning, making them more intuitively accessible. However, while sparsity enhances interpretability, it also presents challenges [72] (Keane et al., 2021). Explanations that are too sparse may suggest statistically or causally weak or irrelevant changes, thereby compromising their informativeness and actionability [74,75] (Byrne, 2019; Keane & Smyth, 2020). For instance, suggesting that changing a single non-causal feature like a zip code would reverse a loan denial may raise questions about the model's fairness or validity. Therefore, while sparsity is desirable for counterfactuals, it must be balanced with plausibility and relevance to ensure that the explanation remains cognitively effective and substantively meaningful.

Plausibility

Sometimes the produced counterfactuals may be unrealistic or impossible to change. Thus, Keane and Smyth (2020) [73] describe plausibility as a problem that occurs when "the counterfactuals generated may not be valid data-points in the domain or they may suggest feature-changes that are difficult-to-impossible" (p. 5). An example is if the counterfactual recommends the user to increase their income by an impossible amount (If you earned $1 million, you would get the loan) or radical propositions (If you changed your gender, you would get the loan). Such implausible counterfactuals erode user trust, limit interpretability, and risk misinforming decision-making [71] (Keane et al., 2021). To enhance plausibility, many counterfactual generation methods incorporate constraints drawn from domain knowledge, data distributions, or causal models to ensure that explanations remain meaningful and actionable.

4.2.2. Causal Modelling

Causal modelling explains not only what correlations exist, but why, allowing users to test interventions and simulate "what–if" scenarios using causal graphs or structural models [43,76]. In process industries, where variables interact continuously, causal models support root-cause analysis and operational reliability. In discrete manufacturing, they diagnose production defects and optimise sequences. Causal modelling strengthens interpretability and supports regulatory transparency by making reasoning chains auditable, a key requirement in emerging compliance frameworks such as the EU AI Act [77,78].

4.2.3. Hybrid Neuro-Symbolic Frameworks

Hybrid neuro-symbolic frameworks integrate neural networks with symbolic rule systems to combine predictive performance with explainable reasoning [79–81]. These approaches are well suited to smart manufacturing environments that integrate robotics, sensing, and real-time control. Symbolic components provide interpretable logic checks that domain experts can validate, while neural components handle complex perception tasks. Such systems also support distributed or federated architectures, maintaining interpretability across decentralised networks [82–84].

4.2.4. Natural Language Generation

Natural language generation (NLG) translates model reasoning into accessible textual explanations [85,86]. In industrial settings, NLG can automatically narrate anomaly detections, maintenance alerts, or decision summaries in operator-appropriate language. This improves clarity, reduces error rates, and supports multilingual communication in global operations. NLG also aligns with regulatory requirements for transparent user-facing explanations under the EU AI Act [87–91].

4.2.5. Visual Analytics Tools

Visual analytics use interactive dashboards, saliency maps, and feature heatmaps to illustrate how models behave [92–97]. In process industries, they highlight relationships among continuous variables; in discrete manufacturing, they support real-time defect detection and maintenance prediction. These tools enhance cross-team interpretability and facilitate auditability, which is increasingly important for regulatory and internal transparency reviews [65,98].

4.2.6. White-Box Modelling

White-box models (e.g., decision trees, generalised additive models) provide direct interpretability and are preferred in high-regulation contexts where audit trails are critical (Figure 11) [99–102]. While they may trade some predictive accuracy for transparency, their simulatability supports operator trust and compliance documentation, aligning with accountability requirements in emerging regulatory regimes [60,65,66,103,104].

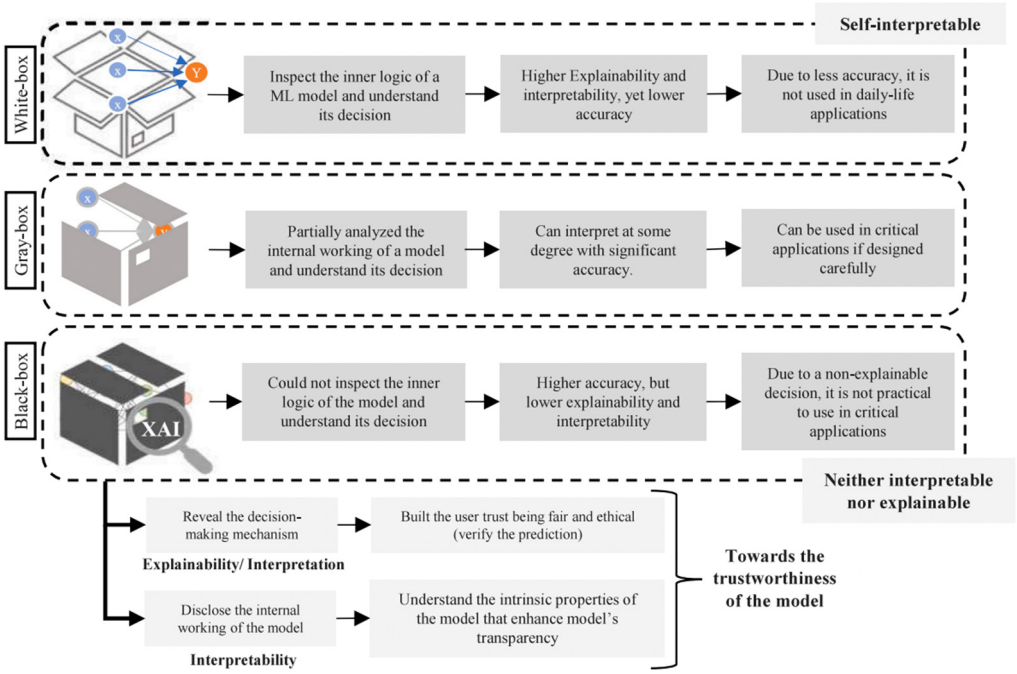

Figure 11. A comparison of white-box, grey-box, and black-box models [99] (Ali et al., 2023).

4.2.7. Interpretable Learning Architectures

Interpretable architectures embed explanation directly into model design through components such as attention mechanisms, prototype networks, or modular layers [102–109]. In process industries, attention layers highlight the most influential operational parameters; in discrete manufacturing, prototype-based classifiers enable explainable visual inspection. These models support measurable improvements in decision confidence and

cognitive workload and can be deployed in distributed industrial environments where local interpretability is required.

In summary, the comparative analysis of XAI techniques demonstrates that no single method universally optimises explainability across all industrial contexts. Instead, their effectiveness depends on the nature of the process, degree of automation, regulatory environment, and human interaction level. By integrating human-centric metrics, regulatory linkages, and decentralised challenges, this revised section highlights how explainable AI can simultaneously enhance operational reliability, ethical accountability, and measurable user trust within both Industry 4.0 and Industry 5.0 paradigms.

5. Conclusions

The accelerating deployment of artificial intelligence (AI) and machine learning across high-impact industrial sectors has generated both unprecedented opportunities and profound challenges, particularly regarding the opacity and accountability of automated decision-making processes. This study sought to address these concerns by systematically reviewing the evolution of Explainable Artificial Intelligence (XAI) within the context of Industry 4.0 and the emerging Industry 5.0 paradigm.

Rather than treating explainability as a purely technical attribute, this review re-frames it as a multidimensional construct encompassing ethical, regulatory, and human-centric dimensions that underpin trustworthy AI. By combining a systematic bibliometric analysis (LRSB) with conceptual synthesis, the study identifies and interprets key research trends, methodological patterns, and theoretical tensions in XAI literature.

The findings indicate a growing convergence between technological transparency and societal accountability. Explainability has evolved from a technical feature into an enabler of ethical governance and human–machine collaboration, with XAI now positioned as a foundation for industrial systems that are interpretable, auditable, and aligned with the principles of the EU AI Act. Techniques such as causal modelling, counterfactual reasoning, and hybrid neuro-symbolic frameworks emerge as crucial pathways toward transparent machine reasoning, fostering greater user trust and organisational legitimacy.

From a theoretical standpoint, this study contributes by articulating a conceptual framework of Transparent Machine Reasoning (TMR), which integrates interpretability, ethical accountability, and regulatory compliance. Practically, it provides a roadmap for applying explainability principles in industrial environments, linking transparency mechanisms to measurable indicators of trust, reliability, and human oversight.

The thematic clusters identified in the literature review directly inform the architecture of the TMR framework. The cluster grounded in explainability principles shapes the normative foundation of the model, while the cluster centred on human-AI interaction informs its emphasis on cognitive alignment and user-oriented interpretability. In turn, the cluster focused on industrial implementation and governance underpins the framework's operational dimension, demonstrating how explainability practices must be embedded within organisational processes, regulatory expectations, and domain-specific constraints. As such, the TMR framework emerges not only as a conceptual contribution but also as an empirically aligned structure capable of guiding the design and evaluation of explainable AI systems in industrial contexts.

As with any systematic literature review, this study was not without limitations. The first concerned the exclusive reliance on the Scopus database, which, although chosen for its breadth, consistency, and robust metadata structure, may have introduced publication and language biases, potentially excluding relevant regional or non-English contributions. To mitigate this limitation, future research could expand the scope of database selection and triangulate searches with complementary indexing platforms such as Web of Science

and IEEE Xplore, thereby increasing representativeness and reducing potential indexing bias. Secondly, the analysis was restricted to peer-reviewed academic sources, thus omitting grey literature, white papers, and preprints that could have provided emerging or practice-oriented perspectives on XAI implementation. Thirdly, while bibliometric tools such as VOSviewer and Bibliometrix offered quantitative rigour, they inherently privileged frequency and co-occurrence patterns over qualitative nuance, meaning that certain contextual or interpretive dimensions may not have been fully captured.

Moreover, the study's temporal scope (up to June 2025) limits its ability to account for rapidly evolving policy frameworks, such as ongoing updates to the EU AI Act and national AI strategies. Lastly, as this research synthesises existing knowledge, it does not include empirical validation or statistical testing of causal relationships among variables, which would be necessary to operationalize the conceptual framework in applied contexts.

Building on these limitations, several avenues for future research are proposed:

- Multi-Database Integration: Future systematic reviews should triangulate Scopus with databases such as Web of Science, IEEE Xplore, or Google Scholar to enhance coverage and reduce indexing bias.
- Longitudinal and Comparative Analyses: Investigations could track how XAI discourse evolves over time and across industrial sectors, mapping the diffusion of explainability practices between manufacturing, healthcare, finance, and logistics.
- Empirical Validation of the TMR Framework: Quantitative studies could operationalize and test the proposed Transparent Machine Reasoning framework through surveys, case studies, or mixed-method approaches.
- Cross-Cultural and Ethical Perspectives: Future work should examine how cultural, institutional, and legal contexts shape perceptions of explainability, fairness, and accountability in AI systems.
- Human-AI Interaction Metrics: Experimental research could measure the behavioural and cognitive effects of explainability tools (e.g., dashboards, visual analytics, or natural language explanations) on user trust, decision accuracy, and satisfaction.
- Policy and Governance Studies: Further exploration is needed into how explainability mechanisms can inform regulatory compliance and organisational governance, particularly within the broader transition toward Industry 5.0.

In conclusion, this study underscores that explainability must be reconceptualized as more than a technical function—it is a strategic and ethical capability essential for the sustainability, accountability, and societal legitimacy of industrial AI. As Industry 5.0 advances, the ability to design systems that are both intelligent and transparent will define the next frontier of responsible technological innovation.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/info16121044/s1.

Author Contributions: Conceptualization, A.T.R. and J.C.D.; methodology, A.T.R. and J.C.D.; software, A.T.R. and J.C.D.; validation, A.T.R. and J.C.D.; formal analysis, A.T.R. and J.C.D.; investigation, A.T.R. and J.C.D.; resources, A.T.R. and J.C.D.; data curation, A.T.R. and J.C.D.; writing—original draft preparation, A.T.R. and J.C.D.; writing—review and editing, A.T.R. and J.C.D.; visualisation, A.T.R. and J.C.D.; supervision, A.T.R. and J.C.D.; project administration, A.T.R. and J.C.D.; funding acquisition, A.T.R. and J.C.D. All authors have read and agreed to the published version of the manuscript.

Funding: The first author receives financial support from the Research Unit on Governance, Competitive-ness and Public Policies (UIDB/04058/2020) + (UIDP/04058/2020), funded by national funds through FCT—Fundação para a Ciência e a Tecnologia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request. https://susy.mdpi.com/user/manuscripts/displayFile/370357210bbe42726c68cc4a22ed82b5/dataset_file (accessed on 5 November 2025).

Acknowledgments: We would like to express our gratitude to the Editor and the Referees. They offered extremely valuable suggestions or improvements. The authors were supported by the GOVCOPP Research Unit of Universidade de Aveiro, Instituto Politécnico de Setúbal, Escola Superior de Ciências Empresariais de Setúbal, Centro de Investigação em Organizações, Mercados e Gestão Industrial (COMEGI), Universidade Lusíada.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A

Table A1. Overview of document citations period ≤2015 to 2025.

Documents		≤2015	2016	2017	2018	2019	2021	2021	2022	2023	2024	2025	Total
Counterfactual explanations for remaining useful life estimation within a Bayesian framework	2025	0	0	0	0	0	0	0	0	0	0	1	1
An extensive bibliometric analysis of artificial intelligence techniques from 2013 to 2023	2025	0	0	0	0	0	0	0	0	0	0	1	1
Generative AI in AI-Based Digital Twins for Fault Diagnosis for Predictive Maintenance in Industry 4.0/5.0	2025	0	0	0	0	0	0	0	0	0	0	3	3
Enhancing transparency and trust in AI-powered manufacturing: A survey of explainable AI (XAI) applications in smart manufacturing in the era of industry 4.0/5.0	2025	0	0	0	0	0	0	0	0	0	0	2	3
How do ML practitioners perceive explainability? an interview study of practices and challenges	2025	0	0	0	0	0	0	0	0	0	0	1	1
A cognitive digital twin for process chain anomaly detection and bottleneck analysis	2025	0	0	0	0	0	0	0	0	0	0	1	1
A Comparative Analysis of LIME and SHAP Interpreters With Explainable ML-Based Diabetes Predictions	2025	0	0	0	0	0	0	0	0	0	6	11	17
Artificial Intelligence and Smart Technologies in Safety Management: A Comprehensive Analysis Across Multiple Industries	2024	0	0	0	0	0	0	0	0	0	0	6	6
Explainable artificial intelligence to increase transparency for revolutionizing healthcare ecosystem and the road ahead	2024	0	0	0	0	0	0	0	0	0	10	10	20
On the Application of Artificial Intelligence/Machine Learning (AI/ML) in Late-Stage Clinical Development	2024	0	0	0	0	0	0	0	0	0	0	1	1
Evaluative Item-Contrastive Explanations in Rankings	2024	0	0	0	0	0	0	0	0	0	0	1	1
Barriers to adopting artificial intelligence and machine learning technologies in nuclear power	2024	0	0	0	0	0	0	0	0	0	2	3	5
Intelligent decision support systems in construction engineering: An artificial intelligence and machine learning approaches	2024	0	0	0	0	0	0	0	0	0	9	15	24
Explainable artificial intelligence (XAI) in finance: a systematic literature review	2024	0	0	0	0	0	0	0	0	0	2	30	32

Table A1. *Cont.*

Documents		≤2015	2016	2017	2018	2019	2021	2021	2022	2023	2024	2025	Total
Leveraging artificial intelligence for enhanced risk management in banking: A systematic literature review	2024	0	0	0	0	0	0	0	0	0	4	1	5
An Interrogative Survey of Explainable AI in Manufacturing	2024	0	0	0	0	0	0	0	0	0	6	9	15
An explanation framework and method for AI-based text emotion analysis and visualisation	2024	0	0	0	0	0	0	0	0	0	9	11	20
Artificial intelligence in manufacturing: Enabling intelligent, flexible and cost-effective production through AI	2024	0	0	0	0	0	0	0	0	0	1	3	4
Recent trends and advances in machine learning challenges and applications for industry 4.0	2024	0	0	0	0	0	0	0	0	0	0	1	1
Harnessing Deep Learning for Fault Detection in Industry 4.0: A Multimodal Approach	2024	0	0	0	0	0	0	0	0	0	0	1	1
A Conceptual Framework for Predictive Digital Dairy Twins: Integrating Explainable AI and Hybrid Modeling	2024	0	0	0	0	0	0	0	0	0	0	1	1
Advancing Manufacturing Through Artificial Intelligence: Current Landscape, Perspectives, Best Practices, Challenges, and Future Direction	2024	0	0	0	0	0	0	0	0	0	1	5	6
Leveraging Information Flow-Based Fuzzy Cognitive Maps for Interpretable Fault Diagnosis in Industrial Robotics	2024	0	0	0	0	0	0	0	0	0	0	2	2
Potential Technological Advancements in the Future of Process Control and Automation	2024	0	0	0	0	0	0	0	0	0	1	2	3
Explainability of Brain Tumor Classification Based on Region	2024	0	0	0	0	0	0	0	0	0	0	4	4
Explainable AI for Industry 5.0: Vision, Architecture, and Potential Directions	2024	0	0	0	0	0	0	0	0	0	6	9	15
Insightful Clinical Assistance for Anemia Prediction with Data Analysis and Explainable AI	2024	0	0	0	0	0	0	0	0	0	2	1	3
Explainable AI for Cyber-Physical Systems: Issues and Challenges	2024	0	0	0	0	0	0	0	0	0	6	6	12
Impact of social media posts' characteristics on movie performance prior to release: an explainable machine learning approach	2024	0	0	0	0	0	0	0	0	0	0	4	4
Trustworthiness of the AI	2024	0	0	0	0	0	0	0	0	0	0	2	2
Explainable Predictive Maintenance: A Survey of Current Methods, Challenges and Opportunities	2024	0	0	0	0	0	0	0	0	0	10	15	25
Explainable AI for 6G Use Cases: Technical Aspects and Research Challenges	2024	0	0	0	0	0	0	0	0	0	10	18	28
Toward Transparent AI for Neurological Disorders: A Feature Extraction and Relevance Analysis Framework	2024	0	0	0	0	0	0	0	0	0	3	5	8
Translating Image XAI to Multivariate Time Series	2024	0	0	0	0	0	0	0	0	0	1	2	3
Explainable Predictive Maintenance of Rotating Machines Using LIME, SHAP, PDP, ICE	2024	0	0	0	0	0	0	0	0	0	9	20	29

Table A1. *Cont.*

Documents		≤2015	2016	2017	2018	2019	2021	2021	2022	2023	2024	2025	Total
AI-Based Task Classification with Pressure Insoles for Occupational Safety	2024	0	0	0	0	0	0	0	0	0	3	2	5
Moving Towards Explainable Artificial Intelligence Using Fuzzy Rule-Based Networks in Decision-Making Process	2024	0	0	0	0	0	0	0	0	0	1	0	1
Deep Learning in Industry 4.0: Transforming Manufacturing Through Data-Driven Innovation	2024	0	0	0	0	0	0	0	0	0	5	7	12
Expl(AI)ned: The Impact of Explainable Artificial Intelligence on Users' Information Processing	2023	0	0	0	0	0	0	0	0	5	37	39	81
ENIGMA: An explainable digital twin security solution for cyber–physical systems	2023	0	0	0	0	0	0	0	0	3	22	21	46
TEA-EKHO-IDS: An intrusion detection system for industrial CPS with trustworthy explainable AI and enhanced krill herd optimization	2023	0	0	0	0	0	0	0	0	2	15	5	22
An optimized model for network intrusion detection systems in industry 4.0 using XAI based Bi-LSTM framework	2023	0	0	0	0	0	0	0	0	6	29	22	57
An Explainable AI Framework for Artificial Intelligence of Medical Things	2023	0	0	0	0	0	0	0	0	0	4	4	8
Explainable AI for Breast Cancer Detection: A LIME-Driven Approach	2023	0	0	0	0	0	0	0	0	0	3	0	3
Explainable Artificial Intelligence: A Study of Current State-of-the-Art Techniques for Making ML Models Interpretable and Transparent	2023	0	0	0	0	0	0	0	0	0	1	1	2
AI-enabled IoT Applications: Towards a Transparent Governance Framework	2023	0	0	0	0	0	0	0	0	0	3	1	4
Towards Explainable AI Validation in Industry 4.0: A Fuzzy Cognitive Map-based Evaluation Framework for Assessing Business Value	2023	0	0	0	0	0	0	0	0	0	0	1	1
A novel Explainable Artificial Intelligence and secure Artificial Intelligence asset sharing platform for the manufacturing industry	2023	0	0	0	0	0	0	0	0	0	3	0	3
XAI Requirements in Smart Production Processes: A Case Study	2023	0	0	0	0	0	0	0	0	0	6	4	10
Introduction to artificial intelligence and current trends	2023	0	0	0	0	0	0	0	0	0	4	4	8
Explainable Artificial Intelligence (XAI) for the Prediction of Diabetes Management: An Ensemble Approach	2023	0	0	0	0	0	0	0	0	2	4	3	9
Explainable Artificial Intelligence (XAI) Based Analysis of Stress Among Tech Workers Amidst COVID-19 Pandemic	2023	0	0	0	0	0	0	0	0	1	7	0	8
A Review of Trustworthy and Explainable Artificial Intelligence (XAI)	2023	0	0	0	0	0	0	0	0	5	61	48	114
6G-BRICKS: Developing a Modern Experimentation Facility for Validation, Testing and Showcasing of 6G Breakthrough Technologies and Devices	2023	0	0	0	0	0	0	0	0	0	2	0	2
Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods	2023	0	0	0	0	0	0	0	0	1	54	41	96
Transparent Artificial Intelligence and Human Resource Management: A Systematic Literature Review	2023	0	0	0	0	0	0	0	0	0	1	0	1

Table A1. *Cont.*

Documents		≤2015	2016	2017	2018	2019	2021	2021	2022	2023	2024	2025	Total
CHAIKMAT 4.0—Commonsense Knowledge and Hybrid Artificial Intelligence for Trusted Flexible Manufacturing	2023	0	0	0	0	0	0	0	0	2	1	1	4
Human-in-Loop: A Review of Smart Manufacturing Deployments	2023	0	0	0	0	0	0	0	0	2	15	4	21
Towards big industrial data mining through explainable automated machine learning	2022	0	0	0	0	0	0	0	8	10	18	12	48
Internet-of-Explainable-Digital-Twins: A Case Study of Versatile Corn Production Ecosystem	2022	0	0	0	0	0	0	0	1	1	3	2	7
Using an Explainable Machine Learning Approach to Minimize Opportunistic Maintenance Interventions	2022	0	0	0	0	0	0	0	0	1	3	1	5
Resource Reservation in Sliced Networks: An Explainable Artificial Intelligence (XAI) Approach	2022	0	0	0	0	0	0	0	0	2	10	5	17
Information Model to Advance Explainable AI-Based Decision Support Systems in Manufacturing System Design	2022	0	0	0	0	0	0	0	1	1	4	3	9
On the Intersection of Explainable and Reliable AI for Physical Fatigue Prediction	2022	0	0	0	0	0	0	0	0	2	10	2	14
Explainable AI for Industry 4.0: Semantic Representation of Deep Learning Models	2022	0	0	0	0	0	0	0	2	7	12	5	26
XAI for operations in the process industry—Applications, theses, and research directions	2021	0	0	0	0	0	0	0	1	0	1	1	3
IEC 61499 Device Management Model through the lenses of RMAS	2021	0	0	0	0	0	0	1	3	0	5	1	10
A human cyber physical system framework for operator 4.0—artificial intelligence symbiosis	2020	0	0	0	0	0	0	10	15	12	15	5	57
Choose for AI and for explainability	2020	0	0	0	0	0	0	0	0	1	0	0	1
Depicting Decision-Making: A Type-2 Fuzzy Logic Based Explainable Artificial Intelligence System for Goal-Driven Simulation in the Workforce Allocation Domain	2019	0	0	0	0	0	0	4	3	2	5	1	15
A fuzzy linguistic supported framework to increase Artificial Intelligence intelligibility for subject matter experts	2019	0	0	0	0	0	0	0	0	1	1	0	2
Interval type-2 fuzzy logic based stacked autoencoder deep neural network for generating explainable AI models in workforce optimization	2018	0	0	0	1	1	2	0	3	2	1	0	10
Working with beliefs: AI transparency in the enterprise	2018	0	0	0	0	2	2	3	9	3	0	0	19
	Total	0	0	0	1	3	4	18	46	74	467	454	1068

References

1. Felzmann, H.; Fosch-Villaronga, E.; Lutz, C.; Tamò-Larrieux, A. Towards transparency by design for artificial intelligence. *Sci. Eng. Ethics* **2020**, *26*, 3333–3361. [CrossRef]
2. Greene, D.; Hoffmann, A.L.; Stark, L. Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. In Proceedings of the 52nd Hawaii International Conference on System Sciences 2019, Maui, HI, USA, 8–11 January 2019. [CrossRef]
3. Hollanek, T. AI transparency: A matter of reconciling design with critique. *Ai Soc.* **2023**, *38*, 2071–2079. [CrossRef]
4. Das, A.; Rad, P. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv* **2020**, arXiv:2006.11371. Available online: https://arxiv.org/pdf/2006.11371 (accessed on 30 June 2025). [CrossRef]

5. Chesterman, S. Through a glass, darkly: Artificial intelligence and the problem of opacity. *Am. J. Comp. Law.* **2021**, *69*, 271–294. [CrossRef]
6. Baum, D.; Baum, K.; Gros, T.P.; Wolf, V. XAI Requirements in Smart Production Processes: A Case Study. In *World Conference on Explainable Artificial Intelligence*; Springer Nature: Cham, Switzerland, 2023; pp. 3–24.
7. Marzi, G.; Balzano, M.; Caputo, A.; Pellegrini, M.M. Guidelines for Bibliometric-Systematic Literature Reviews: 10 steps to combine analysis, synthesis and theory development. *Int. J. Manag. Rev.* **2025**, *27*, 81–103. [CrossRef]
8. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef]
9. Haddaway, N.R.; Page, M.J.; Pritchard, C.C.; McGuinness, L.A. PRISMA 2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimized digital transparency and Open Synthesis. *Campbell Syst. Rev.* **2022**, *18*, e1230. [CrossRef]
10. Rosário, A.T.; Dias, J.C. The New Digital Economy and Sustainability: Challenges and Opportunities. *Sustainability* **2023**, *15*, 10902. [CrossRef]
11. Rosário, A.T.; Raimundo, R. Sustainable Entrepreneurship Education: A Systematic Bibliometric Literature Review. *Sustainability* **2024**, *16*, 784. [CrossRef]
12. Rosário, A.T.; Lopes, P.; Rosário, F.S. Sustainability and the Circular Economy Business Development. *Sustainability* **2024**, *16*, 6092. [CrossRef]
13. Facchini, A.; Termine, A. Towards a taxonomy for the opacity of AI systems. In *Conference on Philosophy and Theory of Artificial Intelligence*; Springer International Publishing: Cham, Switzerland, 2021; pp. 73–89.
14. Chamola, V.; Hassija, V.; Sulthana, A.R.; Ghosh, D.; Dhingra, D.; Sikdar, B. A Review of Trustworthy and Explainable Artificial Intelligence (XAI). *IEEE Access* **2023**, *11*, 78994–79015. [CrossRef]
15. Ganguly, R.; Singh, D. Explainable Artificial Intelligence (XAI) for the Prediction of Diabetes Management: An Ensemble Approach. *Int. J. Adv. Comput. Sci. Appl.* **2023**, *14*, 158–163. [CrossRef]
16. Miltiadou, D.; Perakis, K.; Sesana, M.; Calabresi, M.; Lampathaki, F.; Biliri, E. A novel Explainable Artificial Intelligence and secure Artificial Intelligence asset sharing platform for the manufacturing industry. In Proceedings of the 29th International Conference on Engineering, Technology, and Innovation: Shaping the Future, ICE, Edinburgh, UK, 19–22 June 2023.
17. Salloum, S.A. Trustworthiness of the AI. In *Studies in Big Data*; Springer Science and Business Media: Berlin/Heidelberg, Germany, 2024; Volume 144, pp. 643–650. [CrossRef]
18. Nikiforidis, K.; Kyrtsoglou, A.; Vafeiadis, T.; Kotsiopoulos, T.; Nizamis, A.; Ioannidis, D.; Sarigiannidis, P. Enhancing transparency and trust in AI-powered manufacturing: A survey of explainable AI (XAI) applications in smart manufacturing in the era of industry 4.0/5.0. *ICT Express* **2025**, *11*, 135–148. [CrossRef]
19. Alamgir Kabir, M.; Islam, M.M.M.; Chakraborty, N.R.; Noori, S.R.H. *Trustworthy Artificial Intelligence for Industrial Operations and Manufacturing: Principles and Challenges*; Springer Series in Advanced Manufacturing; Springer Nature: Berlin/Heidelberg, Germany, 2025; Volume Part F138, pp. 179–197. [CrossRef]
20. Benguessoum, K.; Lourenço, R.; Bourel, V.; Kubler, S. *Through the Lens of Explainability: Enhancing Trust in Remaining Useful Life Prognosis Models*; Lecture Notes in Mechanical Engineering; Springer Nature: Cham, Switzerland, 2024; pp. 83–90.
21. Bhattacharya, P.; Obaidat, M.S.; Sanghavi, S.; Sakariya, V.; Tanwar, S.; Hsiao, K.F. Internet-of-Explainable-Digital-Twins: A Case Study of Versatile Corn Production Ecosystem. In Proceedings of the 2022 IEEE International Conference on Communications, Computing, Cybersecurity and Informatics, CCCI, Dalian, China, 17–19 October 2022.
22. Bonci, A.; Longhi, S.; Pirani, M. IEC 61499 Device Management Model through the lenses of RMAS. *Procedia Comput. Sci.* **2021**, *180*, 656–665. [CrossRef]
23. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
24. Bernabé-Moreno, J.; Wildberger, K. A fuzzy linguistic supported framework to increase Artificial Intelligence intelligibility for subject matter experts. *Procedia Comput. Sci.* **2019**, *162*, 865–872. [CrossRef]
25. Bousdekis, A.; Apostolou, D.; Mentzas, G. A human cyber physical system framework for operator 4.0—Artificial intelligence symbiosis. *Manuf. Lett.* **2020**, *25*, 10–15. [CrossRef]
26. Bhattacharya, M.; Penica, M.; O'Connell, E.; Southern, M.; Hayes, M. Human-in-Loop: A Review of Smart Manufacturing Deployments. *Systems* **2023**, *11*, 35. [CrossRef]
27. Singh, L.K.; Khanna, M. Introduction to artificial intelligence and current trends. In *Innovations in Artificial Intelligence and Human-Computer Interaction in the Digital Era*; Elsevier: Amsterdam, The Netherlands, 2023; pp. 31–66. [CrossRef]

28. Arora, N.; Gajjar, Y. Evolution, Need, and Application of Explainable AI in Supply Chain Management. In *Explainable AI and Blockchain for Secure and Agile Supply Chains: Enhancing Transparency, Traceability, and Accountability*; CRC Press: Boca Raton, FL, USA, 2025; pp. 16–35. [CrossRef]
29. Beshaw, F.G.; Atyia, T.H.; Salleh, M.F.M.; Ishak, M.K.; Din, A.S. Utilizing Machine Learning and SHAP Values for Improved and Transparent Energy Usage Predictions. *Comput. Mater. Contin.* **2025**, *83*, 3553–3583. [CrossRef]
30. Alexander, Z.; Chau, D.H.; Saldana, C. An Interrogative Survey of Explainable AI in Manufacturing. *IEEE Trans. Ind. Inform.* **2024**, *20*, 7069–7081. [CrossRef]
31. Chander, A.; Srinivasan, R.; Chelian, S.; Wang, J.; Uchino, K. Working with beliefs: AI transparency in the enterprise. In Proceedings of the IUI Workshops, Tokyo, Japan, 7–11 March 2018.
32. Bajpai, A.; Yadav, S.; Nagwani, N.K. An extensive bibliometric analysis of artificial intelligence techniques from 2013 to 2023. *J. Supercomput.* **2025**, *81*, 540. [CrossRef]
33. Fares, N.Y.; Nedeljkovic, D.; Jammal, M. AI-enabled IoT Applications: Towards a Transparent Governance Framework. In Proceedings of the 2023 IEEE Global Conference on Artificial Intelligence and Internet of Things, GCAIoT, Dubai, United Arab Emirates, 10–11 December 2023.
34. Roy, S.; Pal, D.; Meena, T. Explainable artificial intelligence to increase transparency for revolutionizing healthcare ecosystem and the road ahead. *Netw. Model. Anal. Health Inform. Bioinform.* **2024**, *13*, 4. [CrossRef]
35. Thakur, A.; Vashisth, R.; Tripathi, S. Explainable Artificial Intelligence: A Study of Current State-of-the-Art Techniques for Making ML Models Interpretable and Transparent. In Proceedings of the International Conference on Technological Advancements in Computational Sciences, ICTACS, Tashkent, Uzbekistan, 1–3 November 2023.
36. Votto, A.M.; Liu, C.Z. Transparent Artificial Intelligence and Human Resource Management: A Systematic Literature Review. In Proceedings of the Annual Hawaii International Conference on System Sciences, Online, 3–7 January 2023.
37. Woodbright, M.D.; Morshed, A.; Browne, M.; Ray, B.; Moore, S. Toward Transparent AI for Neurological Disorders: A Feature Extraction and Relevance Analysis Framework. *IEEE Access* **2024**, *12*, 37731–37743. [CrossRef]
38. Byrne, A. Pricing Risk: An XAI Analysis of Irish Car Insurance Premiums. In *World Conference on Explainable Artificial Intelligence*; Springer Nature: Cham, Switzerland, 2024.
39. Park, J.; Kang, D. Artificial Intelligence and Smart Technologies in Safety Management: A Comprehensive Analysis Across Multiple Industries. *Appl. Sci.* **2024**, *14*, 11934. [CrossRef]
40. O'Sullivan, P.; Menolotto, M.; Visentin, A.; O'Flynn, B.; Komaris, D.S. AI-Based Task Classification with Pressure Insoles for Occupational Safety. *IEEE Access* **2024**, *12*, 21347–21357. [CrossRef]
41. Carloni, G.; Berti, A.; Colantonio, S. The role of causality in explainable artificial intelligence. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2025**, *15*, e70015. [CrossRef]
42. Beckers, S. Causal explanations and XAI. In Proceedings of the Conference on Causal Learning and Reasoning, Eureka, CA, USA, 11–13 April 2022. Available online: https://proceedings.mlr.press/v177/beckers22a/beckers22a.pdf (accessed on 30 June 2025).
43. Warren, G.; Keane, M.T.; Byrne, R.M. Features of Explainability: How users understand counterfactual and causal explanations for categorical and continuous features in XAI. *arXiv* **2022**, arXiv:2204.10152. Available online: https://arxiv.org/pdf/2204.10152 (accessed on 30 June 2025). [CrossRef]
44. Černevičienė, J.; Kabašinskas, A. Explainable artificial intelligence (XAI) in finance: A systematic literature review. *Artif. Intell. Rev.* **2024**, *57*, 8–216. [CrossRef]
45. Casalicchio, E.; Gaudenzi, P.; Mancini, L.V. CriSEs: Cybersecurity for small-Satellite Ecosystem-state-of-the-art and open challenge. In Proceedings of the International Astronautical Congress, IAC, Dubai, United Arab Emirates, 12–14 October 2020.
46. Castelnovo, A.; Crupi, R.; Mombelli, N.; Nanino, G.; Regoli, D. Evaluative Item-Contrastive Explanations in Rankings. *Cogn. Comput.* **2024**, *16*, 3035–3050. [CrossRef]
47. Andringa, J.; Baptista, M.L.; Santos, B.F. Counterfactual explanations for remaining useful life estimation within a Bayesian framework. *Inf. Fusion.* **2025**, *118*, 102972. [CrossRef]
48. Cochran, D.S.; Smith, J.; Mark, B.G.; Rauch, E. Information Model to Advance Explainable AI-Based Decision Support Systems in Manufacturing System Design. In *International Symposium on Industrial Engineering and Automation*; Lecture Notes in Networks and Systems; Springer International Publishing: Cham, Switzerland, 2022.
49. Kalmykov, V.L.; Kalmykov, L.V. Towards eXplicitly eXplainable Artificial Intelligence. *Inf. Fusion.* **2025**, *123*, 103352. [CrossRef]
50. Lourenço, A.; Fernandes, M.; Canito, A.; Almeida, A.; Marreiros, G. Using an Explainable Machine Learning Approach to Minimize Opportunistic Maintenance Interventions. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*; Springer International Publishing: Cham, Switzerland, 2022.
51. Banerjee, J.S.; Chakraborty, A.; Mahmud, M.; Kar, U.; Lahby, M.; Saha, G. Explainable Artificial Intelligence (XAI) Based Analysis of Stress Among Tech Workers Amidst COVID-19 Pandemic. In *Advanced AI and Internet of Health Things for Combating Pandemics*; Springer International Publishing: Cham, Switzerland, 2023. [CrossRef]

52. Catti, P.; Bakopoulos, E.; Stipankov, A.; Cardona, N.; Nikolakis, N.; Alexopoulos, K. Human-Centric Proactive Quality Control in Industry 5.0: The Critical Role of Explainable AI. In Proceedings of the 30th ICE IEEE/ITMC Conference on Engineering, Technology, and Innovation: Digital Transformation on Engineering, Technology and Innovation, ICE, Funchal, Portugal, 24–28 June 2024.
53. Tronchin, L.; Cordelli, E.; Celsi, L.R.; MacCagnola, D.; Natale, M.; Soda, P.; Sicilia, R. Translating Image XAI to Multivariate Time Series. *IEEE Access* **2024**, *12*, 27484–27500. [CrossRef]
54. Chochliouros, I.P.; Vardakas, J.; Ramantas, K.; Pollin, S.; Mayrargue, S.; Ksentini, A.; Nitzold, W.; Rahman, M.A.; O'Meara, J.; Chawla, A.; et al. 6G-BRICKS: Developing a Modern Experimentation Facility for Validation, Testing and Showcasing of 6G Breakthrough Technologies and Devices. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*; Springer Nature: Cham, Switzerland, 2023.
55. González-Sendino, R.; Serrano, E.; Bajo, J.; Novais, P. A review of bias and fairness in artificial intelligence. *Int. J. Interact. Multimed. Artif. Intell.* **2024**, *9*, 5–17. [CrossRef]
56. Barnard, P.; MacAluso, I.; Marchetti, N.; Dasilva, L.A. Resource Reservation in Sliced Networks: An Explainable Artificial Intelligence (XAI) Approach. In Proceedings of the IEEE International Conference on Communications, Seoul, Republic of Korea, 16–20 May 2022.
57. Deck, L.; Schoeffer, J.; De-Arteaga, M.; Kuehl, N. A critical survey on fairness benefits of XAI. XAI in Action: Past, Present, and Future Applications. *arXiv* **2023**, arXiv:2310.13007. [CrossRef]
58. Papanikou, V.; Karidi, D.P.; Pitoura, E.; Panagiotou, E.; Ntoutsi, E. Explanations as Bias Detectors: A Critical Study of Local Post-hoc XAI Methods for Fairness Exploration. *arXiv* **2025**, arXiv:2505.00802. Available online: https://arxiv.org/pdf/2505.00802? (accessed on 30 June 2025). [CrossRef]
59. Cummins, L.; Sommers, A.; Ramezani, S.B.; Mittal, S.; Jabour, J.; Seale, M.; Rahimi, S. Explainable Predictive Maintenance: A Survey of Current Methods, Challenges and Opportunities. *IEEE Access* **2024**, *12*, 57574–57602. [CrossRef]
60. Dewasiri, N.J.; Dharmarathna, D.G.; Choudhary, M. Leveraging artificial intelligence for enhanced risk management in banking: A systematic literature review. In *Artificial Intelligence Enabled Management: An Emerging Economy Perspective*; De Gruyter: Berlin, Germany, 2024; pp. 197–213. [CrossRef]
61. Kalyan Chakravarthi, M.; Pavan Kumar, Y.V.; Pradeep Reddy, G. Potential Technological Advancements in the Future of Process Control and Automation. In Proceedings of the 2024 IEEE Open Conference of Electrical, Electronic and Information Sciences, eStream, Vilnius, Lithuania, 25 April 2024.
62. Sivamohan, S.; Sridhar, S.S. An optimized model for network intrusion detection systems in industry 4.0 using XAI based Bi-LSTM framework. *Neural Comput. Appl.* **2023**, *35*, 11459–11475. [CrossRef] [PubMed]
63. Dayanand Lal, N.; Adnan, M.M.; Sutha Merlin, J.; Ramyasree, K.; Palanivel, R. Intrusion Detection System using Improved Wild Horse Optimizer-Based DenseNet for Cognitive Cyber-Physical System in Industry 4.0. In Proceedings of the International Conference on Distributed Computing and Optimization Techniques, ICDCOT, Bengaluru, India, 15–16 March 2024.
64. Natarajan, G.; Elango, E.; Soman, S.; Bai, S.C.P.A. Leveraging Artificial Intelligence and IoT for Healthcare 5.0: Use Cases, Applications, and Challenges. In *Edge AI for Industry 5.0 and Healthcare 5.0 Applications*; CRC Press: Boca Raton, FL, USA, 2025; pp. 153–177. [CrossRef]
65. Laux, J.; Wachter, S.; Mittelstadt, B. Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regul. Gov.* **2024**, *18*, 3–32. [CrossRef]
66. Dikopoulou, Z.; Lavasa, E.; Perez-Castanos, S.; Monzo, D.; Moustakidis, S. Towards Explainable AI Validation in Industry 4.0: A Fuzzy Cognitive Map-based Evaluation Framework for Assessing Business Value. In Proceedings of the 29th International Conference on Engineering, Technology, and Innovation: Shaping the Future, ICE, Edinburgh, UK, 19–22 June 2023.
67. Moorthy, U.M.K.; Muthukumaran, A.M.J.; Kaliyaperumal, V.; Jayakumar, S.; Vijayaraghavan, K.A. Explainability and Regulatory Compliance in Healthcare: Bridging the Gap for Ethical XAI Implementation. *Explain. Artif. Intell. Healthc. Ind.* **2025**, 521–561. [CrossRef]
68. Sonani, R. Hybrid XAI Framework with Regulatory Alignment Metric for Adaptive Compliance Enforcement by Government in Financial Systems. *Acad. Nexus J.* **2024**, *3*. Available online: https://academianexusjournal.com/index.php/anj/article/view/20 (accessed on 30 June 2025).
69. Puthanveettil Madathil, A.; Luo, X.; Liu, Q.; Walker, C.; Madarkar, R.; Qin, Y. A review of explainable artificial intelligence in smart manufacturing. *Int. J. Prod. Res.* **2025**, 1–44. [CrossRef]
70. Kotriwala, A.; Kloepper, B.; Dix, M.; Gopalakrishnan, G.; Ziobro, D.; Potschka, A. XAI for operations in the process industry-Applications, theses, and research directions. In Proceedings of the AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering, Virtual Event, 22–24 March 2021.
71. Dai, X.; Keane, M.T.; Shalloo, L.; Ruelle, E.; Byrne, R.M. Counterfactual explanations for prediction and diagnosis in XAI. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, Oxford, UK, 1–3 August 2022.

72. Keane, M.T.; Kenny, E.M.; Delaney, E.; Smyth, B. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques. *arXiv* **2021**, arXiv:2103.01035. Available online: https://arxiv.org/pdf/2103.01035 (accessed on 30 June 2025). [CrossRef]
73. Keane, M.T.; Smyth, B. Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI (XAI). In *Case-Based Reasoning Research and Development: 28th International Conference, ICCBR 2020, Salamanca, Spain, 8–12 June 2020, Proceedings 28*; Springer International Publishing: Cham, Switzerland, 2020. Available online: https://arxiv.org/pdf/2005.13997 (accessed on 30 June 2025).
74. Bauer, K.; von Zahn, M.; Hinz, O. Expl(AI)ned: The Impact of Explainable Artificial Intelligence on Users' Information Processing. *Inf. Syst. Res.* **2023**, *34*, 1582–1602. [CrossRef]
75. Byrne, R.M. Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019. [CrossRef]
76. Piccininni, M.; Konigorski, S.; Rohmann, J.L.; Kurth, T. Directed acyclic graphs and causal thinking in clinical risk prediction modeling. *BMC Med. Res. Methodol.* **2020**, *20*, 179. [CrossRef]
77. Zhang, Y.; Fitzgibbon, B.; Garofolo, D.; Kota, A.; Papenhausen, E.; Mueller, K. An explainable AI approach to large language model assisted causal model auditing and development. *arXiv* **2023**, arXiv:2312.16211. Available online: https://arxiv.org/pdf/2312.16211 (accessed on 30 June 2025). [CrossRef]
78. Arabikhan, F.; Gegov, A.; Taheri, R.; Akbari, N.; Bader-Ei-Den, M. *Moving Towards Explainable Artificial Intelligence Using Fuzzy Rule-Based Networks in Decision-Making Process*; Lecture Notes in Mechanical Engineering; Springer Nature: Cham, Switzerland, 2024.
79. Chimatapu, R.; Hagras, H.; Starkey, A.; Owusu, G. Interval type-2 fuzzy logic based stacked autoencoder deep neural network for generating explainable AI models in workforce optimization. In Proceedings of the IEEE International Conference on Fuzzy Systems, Rio de Janeiro, Brazil, 8–13 July 2018.
80. Löhr, T. *Identifying a Trial Population for Clinical Studies on Diabetes Drug Testing with Neural Networks*; Lecture Notes in Informatics (LNI); Gesellschaft für Informatik (GI): Bonn, Germany, 2021.
81. Fumanal-Idocin, J.; Andreu-Perez, J. Ex-Fuzzy: A library for symbolic explainable AI through fuzzy logic programming. *Neurocomputing* **2024**, *599*, 128048. [CrossRef]
82. Sarkar, A.; Naqvi, M.R.; Elmhadhbi, L.; Sormaz, D.; Archimede, B.; Karray, M.H. *CHAIKMAT 4.0—Commonsense Knowledge and Hybrid Artificial Intelligence for Trusted Flexible Manufacturing*; Lecture Notes in Mechanical Engineering; Springer International Publishing: Cham, Switzerland, 2023.
83. Shen, J.; Morrison, M.; Miao, H.; Gu, F. Harnessing Deep Learning for Fault Detection in Industry 4.0: A Multimodal Approach. In Proceedings of the 2024 IEEE 6th International Conference on Cognitive Machine Intelligence (CogMI), Washington, DC, USA, 28–31 October 2024.
84. Wojak-Strzelecka, N.; Bobek, S.; Nalepa, G.J.; Stefanowski, J. Towards Differentiating Between Failures and Domain Shifts in Industrial Data Streams. In Proceedings of the CEUR Workshop Proceedings, Kyiv, Ukraine, 24–27 January 2024.
85. Baaj, I.; Poli, J.P. Natural language generation of explanations of fuzzy inference decisions. In Proceedings of the 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), New Orleans, LA, USA, 23–26 June 2019. [CrossRef]
86. Danilevsky, M.; Qian, K.; Aharonov, R.; Katsis, Y.; Kawas, B.; Sen, P. A survey of the state of explainable AI for natural language processing. *arXiv* **2020**, arXiv:2010.00711. [CrossRef]
87. El-Assady, M. Challenges and Opportunities in Text Generation Explainability. *arXiv* **2024**, arXiv:2405.08468. [CrossRef]
88. Mariotti, E.; Alonso, J.M.; Gatt, A. Towards harnessing natural language generation to explain black-box models. In *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 22–27.
89. Narteni, S.; Orani, V.; Cambiaso, E.; Rucco, M.; Mongelli, M. On the Intersection of Explainable and Reliable AI for Physical Fatigue Prediction. *IEEE Access* **2022**, *10*, 76243–76260. [CrossRef]
90. Li, Y.; Chan, J.; Peko, G.; Sundaram, D. An explanation framework and method for AI-based text emotion analysis and visualisation. *Decis. Support Syst.* **2024**, *178*, 114121. [CrossRef]
91. Mohammadi, A.; Maghsoudi, M. Bridging perspectives on artificial intelligence: A comparative analysis of hopes and concerns in developed and developing countries. *AI Soc.* **2025**, *40*, 5713–5734. [CrossRef]
92. Li, Z.; Ding, Y.; Lei, Y.; Oliveira, F.J.M.S.; Neto, M.J.P.; Kong, M.S.M. Integrating artificial intelligence in industrial design: Evolution, applications, and future prospects. *Int. J. Arts Technol.* **2024**, *15*, 139–169. [CrossRef]
93. Ahmed, S.; Kaiser, M.S.; Shahadat Hossain, M.; Andersson, K. A Comparative Analysis of LIME and SHAP Interpreters with Explainable ML-Based Diabetes Predictions. *IEEE Access* **2025**, *13*, 37370–37388. [CrossRef]
94. Nazim, S.; Alam, M.M.; Rizvi, S.S.; Mustapha, J.C.; Hussain, S.S.; Suud, M.M. Advancing malware imagery classification with explainable deep learning: A state-of-the-art approach using SHAP, LIME and Grad-CAM. *PLoS ONE* **2025**, *20*, e0318542. [CrossRef]

95. Narkhede, J. Comparative Evaluation of Post-Hoc Explainability Methods in AI: LIME, SHAP, and Grad-CAM. In Proceedings of the 2024 4th International Conference on Sustainable Expert Systems (ICSES), Kaski, Nepal, 15–17 October 2024. [CrossRef]
96. Agrawal, K.; Nargund, N. Deep Learning in Industry 4.0: Transforming Manufacturing Through Data-Driven Innovation. In *International Conference on Distributed Computing and Intelligent Technology*; Springer Nature: Cham, Switzerland, 2024.
97. Bhati, D.; Neha, F.; Amiruzzaman, M. A survey on explainable artificial intelligence (xai) techniques for visualizing deep learning models in medical imaging. *J. Imaging* **2024**, *10*, 239. [CrossRef] [PubMed]
98. Aldughayfiq, B.; Ashfaq, F.; Jhanjhi, N.Z.; Humayun, M. Explainable AI for retinoblastoma diagnosis: Interpreting deep learning models with LIME and SHAP. *Diagnostics* **2023**, *13*, 1932. [CrossRef]
99. Ali, S.; Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J.M.; Confalonieri, R.; Guidotti, R.; Del Ser, J.; Díaz-Rodríguez, N.; Herrera, F. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Inf. Fusion* **2023**, *99*, 101805. [CrossRef]
100. Salih, A.M.; Wang, Y. Are Linear Regression Models White Box and Interpretable? *arXiv* **2024**, arXiv:2407.12177. [CrossRef]
101. Moreira, C.; Chou, Y.L.; Hsieh, C.; Ouyang, C.; Pereira, J.; Jorge, J. Benchmarking instance-centric counterfactual algorithms for XAI: From white box to black box. *ACM Comput. Surv.* **2025**, *57*, 1–37. [CrossRef]
102. Abdulrashid, I.; Ahmad, I.S.; Musa, A.; Khalafalla, M. Impact of social media posts' characteristics on movie performance prior to release: An explainable machine learning approach. *Electron. Commer. Res.* **2024**, 1–25. [CrossRef]
103. Garouani, M.; Ahmad, A.; Bouneffa, M.; Hamlich, M.; Bourguin, G.; Lewandowski, A. Towards big industrial data mining through explainable automated machine learning. *Int. J. Adv. Manuf. Technol.* **2022**, *120*, 1169–1188. [CrossRef]
104. Guidotti, D.; Pandolfo, L.; Pulina, L. A Systematic Literature Review of Supervised Machine Learning Techniques for Predictive Maintenance in Industry 4.0. *IEEE Access* **2025**, *13*, 102479–102504. [CrossRef]
105. Waqar, A. Intelligent decision support systems in construction engineering: An artificial intelligence and machine learning approaches. *Expert Syst. Appl.* **2024**, *249*, 123503. [CrossRef]
106. Hall, A.; Agarwal, V. Barriers to adopting artificial intelligence and machine learning technologies in nuclear power. *Prog. Nucl. Energy* **2024**, *175*, 105295. [CrossRef]
107. Köchert, K.; Friede, T.; Kunz, M.; Pang, H.; Zhou, Y.; Rantou, E. On the Application of Artificial Intelligence/Machine Learning (AI/ML) in Late-Stage Clinical Development. *Ther. Innov. Regul. Sci.* **2024**, *58*, 1080–1093. [CrossRef] [PubMed]
108. Rodriguez-Fernandez, V.; Camacho, D. Recent trends and advances in machine learning challenges and applications for industry 4.0. *Expert Syst.* **2024**, *41*, e13506. [CrossRef]
109. Terziyan, V.; Vitko, O. Explainable AI for Industry 4.0: Semantic Representation of Deep Learning Models. *Procedia Comput. Sci.* **2022**, *200*, 216–226. [CrossRef]