Ethical Challenges in Artificial Intelligence: Navigating the Boundaries of Responsible Innovation

1st Julio C. S. Lima
Post-Graduate Program of Systems and Automation Engineering
Federal University of Lavras - UFLA
Lavras/MG, Brazil
lima.julio.cs@gmail.com

2nd Felipe O. Silva
Department of Automatics
Federal University of Lavras - UFLA
Lavras/MG, Brazil
felipe.oliveira@ufla.br

3rd Danton Diego Ferreira
Department of Automatics
Federal University of Lavras - UFLA
Lavras/MG, Brazil
danton@ufla.br

***Abstract*—The integration of Artificial Intelligence (AI) in Industry 4.0 and 5.0 demands robust ethical frameworks; however, the practical application of principles such as fairness, accountability, and transparency in Computational Intelligence (CI) techniques remains a challenge. This study presents a Systematic Literature Review (SLR), following PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) guidelines, analyzing peer-reviewed articles (from 2018 to early 2025) from six databases on how these principles are addressed. The review maps the discourse across eight thematic axes, including algorithmic bias mitigation, explainability (XAI), human-centric AI design, and governance mechanisms. Findings reveal a persistent gap between widely accepted ethical principles and their fragmented implementation in socio-technical systems. Although human-centered approaches and fairness-aware tools continue to advance, effectively operationalizing ethics still requires further development. This review consolidates the current state of research, identifies key challenges facing the CI community, and outlines directions for promoting ethically aligned AI innovation in industrial contexts.**

***Index Terms*—Artificial Intelligence, Ethics, Industry 4.0, Human-Centric AI, Fairness, Accountability, Governance, Algorithmic Bias, Systematic Literature Review**

I. Introduction

The widespread integration of Artificial Intelligence (AI) systems within Industry 4.0 has markedly enhanced automation, predictive analytics, and autonomous decision-making across manufacturing, logistics, and supply chain domains. Among these technologies, deep learning has emerged as a key enabler for high-dimensional perception and complex control tasks, such as quality inspection, demand forecasting, and robotic navigation. However, the inherent opacity of these models — often referred to as "black boxes" — has triggered growing ethical concerns regarding transparency, interpretability, and trustworthiness in AI-driven industrial environments [1]–[3]. This lack of explainability not only impairs human oversight but also raises critical issues related to algorithmic bias, the erosion of accountability, and the risk of disproportionate harm to marginalized or underrepresented communities [4].

Similarly, evolutionary algorithms and swarm intelligence techniques — widely applied in multi-objective optimization problems such as production scheduling, logistics planning, and resource allocation — introduce distinct ethical challenges concerning fairness, traceability, and responsibility. These methods operate through stochastic exploration and adaptive heuristics, which often lack clear interpretability or reproducible reasoning paths. As a result, it becomes difficult to audit outcomes, assess causality, or ensure that optimization objectives do not inadvertently embed or amplify existing inequities [2], [5], [6].

The integration of these AI techniques into cyber-physical production systems, smart factories, and autonomous industrial agents has intensified the need for ethical, legal, and governance frameworks. Decisions made by opaque or non-accountable AI systems may affect worker roles, safety, and access to opportunities — raising concerns about transparency, human oversight, and institutional accountability. These issues are particularly critical in industrial contexts, where operational efficiency often conflicts with fairness and explainability [7], [8].

In this context, this study aims to investigate how key ethical principles — namely fairness, accountability, and transparency — are addressed in recent academic literature. To that end, a Systematic Literature Review (SLR) was conducted, grounded in the PRISMA methodology [9] and structured around eight thematic axes, including AI, ethical principles, governance and responsibility, transparency and explainability, algorithmic bias and fairness, Human-Centric AI (HCAI), regulation and public policy, and industrial applications.

The review is guided by three research questions: (i) What are the key ethical concerns related to the implementation of AI systems? (ii) How are principles such as fairness, accountability, and transparency addressed in current literature? (iii) What frameworks and practices are being proposed to mitigate these ethical risks in Industry 4.0?

By addressing these questions, this study contributes to a theoretical and practical foundation for responsible AI in industrial settings. The answers to RQ1–RQ3 are presented in the Literature Review and Discussion (Section IV) as follows: RQ1 in Subsection IV-A, RQ2 in Subsection IV-B, and RQ3 in Subsection IV-C; a deeper treatment of algorithmic bias and fairness is provided in Subsection IV-D.

II. Methodology

Eligibility Criteria

We included peer-reviewed journal and conference papers published between **2018** and **March 2025**, in English or Portuguese, reporting *methods, frameworks, or empirical evidence* related to fairness, accountability, transparency, human-centric AI, or governance in industrial/organizational contexts. We excluded editorials, position papers lacking method, duplicates, and non-sociotechnical applications.

Search Strategy

Databases: ScienceDirect, Scopus, Web of Science, IEEE Xplore, SpringerLink, and CAPES. Final search date: **March 2025**. Example (Scopus, TITLE-ABS-KEY): *("artificial intelligence" OR AI) AND (ethics OR "ethical principles" OR governance OR accountability OR bias OR fairness OR justice) AND ("human-centric AI" OR "human-centered AI" OR "trustworthy AI" OR XAI OR "responsible AI") AND ("Industry 4.0" OR "cyber-physical systems" OR "smart manufacturing" OR "autonomous systems")*. Strings were adapted per database. Reporting follows **PRISMA 2020** [9].

Screening and Reliability

Two-stage screening (titles/abstracts; full text) by two independent reviewers; disagreements resolved by consensus. Inter-rater reliability on a pilot of **120** records yielded **Cohen's** $\kappa = 0.78$.

Data Extraction and Synthesis

We used a standardized template (metadata, domain, study type, ethical principles, techniques/artifacts, metrics, risks, industrial sector, cited regulations). Synthesis combined *narrative* and *descriptive statistics* (frequencies, temporal trends) with *thematic analysis* aligned to the eight axes.

Quality Appraisal

We applied **MMAT 2018/2022** checklists to assess study quality (high/moderate/low). Low-quality studies were retained for transparency and analyzed in sensitivity checks [51], [52].

Screening Reliability and Data Items: Two independent reviewers screened records; disagreements were resolved by consensus. Pilot reliability: **Cohen's** $\kappa = 0.78$ on **120** records. Data items extracted: venue, sector, study type, principles, techniques/artefacts, metrics, risks, jurisdiction/regulations, and quality (MMAT).

III. Results

A. Study Selection

We identified **1,540** records across databases, removed **420** duplicates, and screened **1,120** titles/abstracts. After full-text assessment of **270** articles, **112** studies met the inclusion criteria. Inter-rater reliability on a pilot sample of **120** records yielded **Cohen's** $\kappa = 0.78$ (substantial agreement).

B. Corpus Characterization

Table I summarizes sectors, study types, regulations cited (non-exclusive), quality appraisal (MMAT), and venue type for the **112** included studies. Percentages are rounded.

TABLE I
CORPUS CHARACTERIZATION SUMMARY (N = 112).

Category	Count	%
Sectors		
Manufacturing	38	33.9
Healthcare	25	22.3
Logistics	18	16.1
Energy/Utilities	11	9.8
Other (finance, public, etc.)	20	17.9
Study types		
Empirical (case/field/lab)	54	48.2
Proposals/Methods	40	35.7
Reviews/Surveys	18	16.1
Regulations/standards cited (non-exclusive)		
GDPR/LGPD	69	61.6
EU AI Act (draft/2024 text)	46	41.1
ISO/IEC standards (e.g., 42001)	30	26.8
Quality appraisal (MMAT)		
High	49	43.8
Moderate	43	38.4
Low	20	17.9
Venue type		
Journal	69	61.6
Conference	43	38.4

TABLE II
THEME FREQUENCIES ACROSS THE EIGHT AXES (NON-EXCLUSIVE; N = 112).

Axis	Examples	Count	%
Transparency / XAI	post-hoc, intrinsic, counterfactuals	65	58.0
Governance / Accountability	policies, logs, audits	60	53.6
Bias / Fairness	metrics, mitigation, audits	55	49.1
Human-Centric AI	HCI/HRI, oversight	52	46.4
Regulation / Policy	GDPR/LGPD, EU AI Act, ISO/IEC	44	39.3
Industrial Applications	CPS, smart manufacturing	39	34.8
Ethical Principles	justice, autonomy, beneficence	37	33.0
Human Oversight	HITL, approval loops	31	27.7

TABLE III
YEARLY DISTRIBUTION OF INCLUDED STUDIES (2018–2025).

Year	2018	2019	2020	2021	2022	2023	2024	2025*
Count	6	10	14	18	20	20	16	8

*Partial year (Jan–Mar).

C. Primary Findings

Thematic prevalence (Tables I–II) confirms a strong emphasis on transparency/XAI and governance/accountability, with fairness/bias as a sustained concern across sectors. Despite growing references to GDPR/LGPD and the EU

AI Act, auditable mechanisms (decision logs, provenance, standardized documentation) remain underreported in production environments. The yearly trend (Table III) peaks in 2022–2023, consistent with accelerated regulatory debate and maturing industrial adoption.

IV. Literature Review

The rapid evolution of AI has significantly transformed various industrial sectors, accelerating the shift from Industry 4.0 to Industry 5.0. This new paradigm emphasizes human centrality, promoting a harmonious integration between advanced technologies and fundamental human values. However, this transformation raises complex ethical concerns related to algorithmic transparency, fairness, responsibility, and worker well-being.

Recent studies have highlighted the need for more human-centered approaches to AI deployment. For instance, the systematic review by Grosse et al. [10] identifies a significant gap in the consideration of psychosocial factors such as trust, autonomy, and worker motivation in Industry 4.0 environments. Similarly, Tahei et al. [11] emphasize that existing research predominantly focuses on governance and justice, calling for an expanded scope that includes privacy, security, and human flourishing.

Khan et al. [12] identify widely acknowledged ethical principles in the literature — such as transparency, privacy, accountability, and fairness — but highlight challenges in the practical implementation of these principles due to a lack of clear guidelines and ethical literacy. Complementing this view, Gao et al. [13] underscore critical dilemmas such as the Collingridge dilemma and the need for robust ethical frameworks to guide AI development.

In the industrial context, Bibby and Haverly [14], and Peres et al. [15] explore the maturity of Industry 4.0 and the integration of cyber-physical systems designed around human operators, respectively, reinforcing the importance of maintaining human agency in automated operations. Moreover, Zhu and Luo [16] propose a framework for artificial empathy in human-centered design, emphasizing the integration of creativity, empathy, and collaboration as foundational principles in the development of AI systems.

This body of evidence suggests that despite meaningful advances in AI deployment across industrial settings, a more holistic approach is still needed — one that effectively integrates ethical reasoning and human-centric values. The following Sections delve deeper into these discussions by analyzing the prevailing frameworks and gaps in the current literature on ethical and human-aligned AI in Industry 4.0 and 5.0 contexts.

A. Ethical Principles and Key Concerns in AI Implementation

Core principles—transparency, fairness, accountability, and privacy—recur across the literature, yet translating them into actionable requirements remains difficult; guidelines risk abstraction or "ethics washing" without enforceable mechanisms [12], [17]. Model opacity is central: deep networks and complex rule bases deliver accuracy with limited insight, motivating XAI to balance intelligibility and performance [2], [18], [19].

A "value translation gap" persists between policy commitments and design practice, especially in high-stakes domains where explainability affects legal accountability, reliability, and safety [7]. Performance-centric cultures can drive ethical fading—deprioritizing fairness and oversight—while opacity complicates rights protection and responsibility assignment [20]–[24].

Moreover, the problem of ethical fading — the gradual erosion of moral awareness in highly technical, goal-oriented environments — is especially pertinent in industrial AI systems [20]. When optimization and performance metrics dominate development priorities, ethical considerations may be deprioritized, allowing decisions to become automated and unchallenged — even when they perpetuate unfair or harmful outcomes.

Binns [21] and Mittelstadt [22] underscore the risks posed by algorithmic opacity. Particularly in systems based on deep learning, AI may become functionally opaque to both users and developers. This lack of interpretability not only undermines trust, but also complicates efforts to determine whether rights have been violated — or to assign responsibility in the event of harm.

Ethical risks are also unevenly distributed. Eubanks [23], D'Ignazio and Klein [24] show that marginalized populations are disproportionately exposed to algorithmic bias, surveillance, and exclusion, reinforcing structural inequalities. These findings underscore the need for intersectional, context-aware ethical evaluations, especially in industrial environments where labor dynamics and automation intersect with social vulnerability.

In summary, while principles such as transparency, accountability, and fairness are widely acknowledged in AI ethics discourse, their consistent and systematic implementation in real-world AI systems — particularly within Industry 4.0 contexts — remains limited and fragmented. The next section explores how these principles are interpreted, enforced, or challenged, with a focus on algorithmic fairness and governance mechanisms.

B. Human-Centric AI and Industry 5.0

The shift from Industry 4.0 to Industry 5.0 represents a significant transition — from automation-centric production toward systems that prioritize human values and well-being. This evolution introduces new demands for Computational Intelligence (CI) techniques, pushing AI beyond efficiency alone. Today's systems must collaborate with humans, adapt to dynamic environments, and align with ethical standards and human values.

This paradigm shift drives research into areas such as safe and intuitive Human-Robot Interaction (HRI), robust machine learning with limited or noisy data, and XAI to support human supervision. The overarching goal is to enhance human capabilities rather than replace them, fostering inclusive and sustainable industrial ecosystems [10], [25], [26].

In response to concerns like algorithmic opacity and the dehumanization of labor in automated environments, Human-Centric Artificial Intelligence (HCAI) has emerged as a guiding concept. It promotes ethical alignment in system design and deployment.

According to Breque et al. [27], and echoed by scholars like Xu et al. [28] and Adel [29], Industry 5.0 emphasizes resilience, sustainability, and human-centricity (Fig. 1). Technological advancement, in this view, must be directed by societal and ethical priorities. Fairness, accountability, and transparency are no longer abstract ideals — they become design imperatives, especially in sensitive sectors such as manufacturing, logistics, and healthcare.

Recent frameworks seek to operationalize these values. The High-Level Expert Group on AI [30] of the European Commission proposed seven key requirements for ethical AI, including human agency, technical robustness, privacy, and non-discrimination. These principles guide both policy and design. Yet, studies by Moley et al. [31] and Fjeld et al. [32] point out that implementation remains fragmented, especially outside the European Union.

In practice, human-centric AI in industrial contexts involves collaboration between humans and machines. For instance, Zuo and Luo [16] advocate for embedding empathy, creativity, and collaboration into AI architectures, particularly in cognitively complex environments with continuous human-machine interaction.

Another critical dimension of human–AI integration is system transparency. While deep learning has become a cornerstone of automation — enabling perception, monitoring, and decision-making — its complexity has raised concerns about opacity, reliability, and trust. As a response, XAI has gained momentum by developing models that are more interpretable, auditable, and understandable.

XAI plays a crucial role in fostering trust and accountability in safety-critical environments [33]–[35]. Tools like LIME, introduced by Ribeiro et al. [36], have become foundational in interpreting predictions from black-box models. However, as Durán and Jongsma [37] cautions, technical explainability must be paired with user-friendly interfaces and clear institutional responsibilities to be effective.

Fairness and accountability, moreover, are not fixed standards but contextual and relational. Whittlestone et al. [38] advocate for moving beyond static ethical checklists, recommending participatory governance that includes workers, designers, and affected communities. This participatory approach allows for adaptive frameworks that reflect local norms, labor conditions, and operational risks.

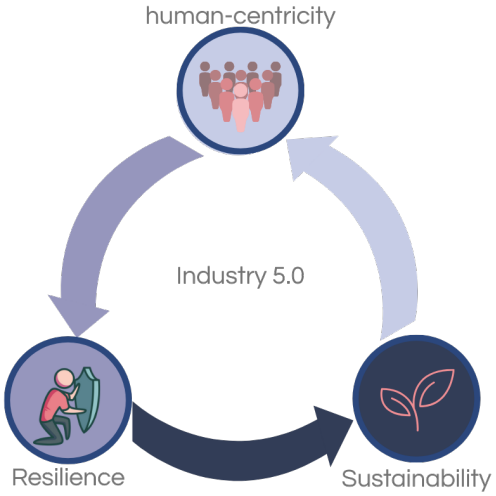

Fig. 1. Core values of Industry 5.0 [27]

In summary, embedding principles like fairness, transparency, and accountability throughout the AI lifecycle — from design and data collection to deployment and monitoring — is essential in the context of Industry 5.0. Meeting this challenge requires not only technical innovation but also institutional change, cultural adaptation, and sustained interdisciplinary collaboration.

C. Governance, Regulation, and Accountability in AI Implementation

As AI systems become increasingly integrated into industrial processes, the governance of ethical risks associated with automation, decision-making, and data usage has become a priority for researchers, regulators, and professionals. Ensuring that AI operates in alignment with societal values — particularly in complex and high-risk environments such as those found in Industry 4.0 — requires the implementation of robust regulatory structures and governance mechanisms that promote accountability, traceability, and continuous oversight.

A key contribution to this regulatory landscape is the European Commission's proposal [39], which introduces a risk-based classification system for AI systems. Under this model, high-risk systems — such as those applied in critical infrastructure, recruitment processes, or biometric identification — must meet stringent requirements regarding transparency, human oversight, and technical robustness. According to Veale and Borgesius [40], this approach represents a regulatory milestone, offering a concrete legal framework where previously only normative guidelines prevailed.

In parallel with legislative efforts, various soft governance tools have emerged, such as ethical guidelines, technical standards, and evaluation frameworks. In a meta-analysis of over 80 AI ethics documents, Fjeld et al. [32] identified recurring normative commitments to fairness, transparency, and accountability. However, the authors also highlight the absence of enforcement mechanisms and the overlap between corporate and governmental discourse, which may undermine the practical effectiveness of such guidelines.

For industrial applications, the Accountability Framework for Industrial AI proposed by Winfield and Jirotka [41] stands out. It advocates for the implementation of "ethical black boxes" — data logging mechanisms embedded in AI systems to enable post hoc audits. These mechanisms are especially relevant in environments where human operators rely on autonomous systems for safety-critical tasks, such as in manufacturing, logistics, or energy networks.

Complementing these frameworks, corporate governance models aim to embed ethical reflection throughout the AI lifecycle. The model proposed by Ashok et al. [42], for example, combines organizational policies, multidisciplinary ethics boards, and ongoing risk audits, forming an enterprise-level ethical governance structure. Such an approach is particularly suitable for large-scale industrial deployments, where regulatory compliance alone may be insufficient to mitigate dynamic and contextual risks.

Despite these advances, several scholars warn of potential pitfalls. For Mittelstadt [22], abstract principles alone do not guarantee ethical outcomes in AI applications; without operational strategies, such principles risk becoming rhetorical tools. Similarly, Dignum [43] emphasizes the importance of co-governance, advocating for the active participation of multiple stakeholders — workers, regulators, designers, and affected communities — in defining and monitoring AI responsibilities.

In the specific context of Industry 4.0, Coelho et al. [44] propose a comprehensive framework for AI governance in organizations, with an emphasis on the creation of ethics committees as a central mechanism for ensuring responsible practices aligned with ethical principles. This practice fosters an organizational culture of accountability among engineers and managers, transforming ethical governance into not only a legal requirement but also a competitive advantage in human-centered industries.

In summary, the current regulatory frameworks and governance models provide an essential foundation for the ethical implementation of AI in Industry 4.0. However, sustainable progress depends on bridging the gap between principle and practice — requiring the incorporation of accountability at both technical and organizational levels, as well as the effective engagement of all actors across the value chain.

1) Legal and Ontological Grounding for AI Governance: In industrial ecosystems, governance should operationalize *duty of care*, accountability and traceability with meaningful human oversight, supported by risk-based assessments (e.g., DPIA under GDPR/LGPD) and auditable controls [40], [62]. We propose an ontological scaffold: {Agent, System, Decision, Evidence, Harm, Affected party, Context, Normative requirement} with relations (*produces*, *justifies*, *impacts*, *is-responsible-for*, *is-governed-by*). Coupled with documentation artifacts—*Model Cards* and *Datasheets*—and management standards (e.g., ISO/IEC 42001), this enables audit queries such as "who decided what, based on which evidence, under which constraints, and why alternative options were rejected?" [53], [54], [61].

D. Algorithmic Bias and Fairness in Socio-Technical Systems

Algorithmic systems are often perceived as objective tools, yet growing empirical evidence reveals that algorithmic bias is a pervasive and systemic issue across domains such as hiring, lending, policing, and healthcare. In industrial and operational contexts, where data-driven decision-making is increasingly central, such biases pose critical ethical, legal, and reputational risks.

Bias in AI systems typically originates from three primary sources: historical or unbalanced training data, flawed model assumptions, and systemic social inequalities that are reproduced in technical infrastructures [45]. For instance, Obermeyer et al. [46] demonstrated how a widely used healthcare algorithm in the U.S. encoded racial bias by underestimating the medical needs of Black patients based on cost-based proxies. These findings underscore the urgent need for auditing practices and fairness-aware machine learning frameworks.

Several taxonomies of fairness have emerged in the literature. Binns [21] applies insights from political philosophy to categorize fairness notions into procedural, distributive, and relational dimensions, emphasizing the importance of contextualizing fairness depending on the system's function and the population affected. Similarly, Barocas et al. [47] argue that technical fairness metrics (e.g., equalized odds, demographic parity) must be interpreted alongside normative and institutional considerations, especially when trade-offs are inevitable.

In industrial contexts, algorithmic bias may appear in areas such as automated recruitment, predictive maintenance, or resource allocation—where decisions affect employment, safety, and workload distribution. Holstein et al. [48] note that the lack of practical tools and organizational incentives for mitigating bias leads many teams to deprioritize fairness, especially in high-pressure environments.

To address these risks, a number of tools and practices have been proposed. One example is IBM's AI Fairness 360 (AIF360) toolkit, which provides open-source libraries for bias detection and mitigation [49]. However, technical interventions alone are insufficient. As D'Ignazio and Klein [24] argue, "data feminism" offers a critical lens to expose how power dynamics shape data collection, labeling, and usage. They advocate for inclusive data practices and participatory design as foundational to ethical AI.

Accountability is also key to fairness. Selbst et al. [50] introduce the concept of "the portability trap", where fairness solutions developed in one context are mistakenly assumed to be transferable across domains. This reinforces the need for localization, stakeholder engagement, and continuous monitoring.

In sum, mitigating algorithmic bias requires a multi-layered approach that encompasses technical robustness, model interpretability, inclusive data governance, and institutional accountability. For Industry 4.0 and 5.0 to evolve along ethical lines, fairness must be embedded not only in the algorithms themselves but also within the broader organizational culture and policy frameworks that guide their use.

To further illustrate these considerations, Table IV presents a summary of key ethical dilemmas associated with artificial intelligence and contrasts them with corresponding principles of responsible innovation. This comparison highlights the practical implications of these issues for the development, deployment, and governance of AI technologies across industrial settings.

Finally, to synthesize the core concepts discussed throughout this section, Fig. 2 illustrates the cyclical process of responsible innovation in artificial intelligence. The diagram emphasizes the interdependent stages — including impact assessment, stakeholder engagement, transparency, accountability, and continuous improvement — that collectively guide the ethical design and implementation of AI systems. This model reinforces the idea that ethical AI is not a static end state but an ongoing commitment throughout the system's lifecycle.

V. Concluding Remarks on the Literature Review

The literature reviewed in this Section demonstrates a growing academic and institutional commitment to understanding and addressing the ethical implications of AI, partic-

TABLE IV
ETHICAL DILEMMAS VS. PRINCIPLES OF RESPONSIBLE INNOVATION

Ethical Dilemmas in AI	**Principles of Responsible Innovation**	**Practical Implications**
Algorithmic discrimination	Justice and Inclusion	Ensure data diversity and bias audits.
Opaqueness and "black-box" decision-making	Transparency and Explainability	Develop interpretable AI and communicate decisions clearly.
Diffused responsibility (who is accountable?)	Accountability	Establish clear technical and legal frameworks for liability.
Automation without human oversight	Human Autonomy and Control	Preserve human-in-the-loop systems in critical decisions.
Exploitation of personal data	Privacy and Informed Consent	Apply privacy-by-design and ensure user awareness and consent.
Profit-driven innovation vs. public good	Sustainability and Collective Interest	Align technological goals with long-term social and environmental values.

ularly in the context of Industry 4.0 and 5.0. Across all axes explored — ranging from foundational ethical principles to applied governance and algorithmic fairness — a common thread emerges: the urgent need for operationalizing ethical values within socio-technical systems.

In response to RQ1, we observed that ethical concerns surrounding AI implementation are multidimensional, encompassing transparency, bias, accountability, privacy, and human dignity. These challenges are especially pronounced in industrial environments, where automated systems interact with human workers and impact decisions with material consequences.

Addressing RQ2, recent literature reveals that although principles such as fairness, accountability, and transparency

Fig. 2. The Responsible Innovation Cycle in Artificial Intelligence. A visual representation of key principles and feedback loops that underpin ethical and sustainable AI development

are widely recognized, their practical application remains inconsistent. Human-centric frameworks are gaining prominence as a corrective response, calling for inclusive design processes, interpretable models, and value-sensitive system development. Notably, Industry 5.0 repositions AI from a tool of automation to a means of enhancing human well-being and creativity.

Finally, in relation to RQ3, multiple regulatory and governance proposals—ranging from the EU AI Act to localized organizational ethics frameworks — are beginning to take shape. However, the literature warns that principles without implementation mechanisms are insufficient. Sustained progress will require institutional accountability, multi-stakeholder participation, and continuous evaluation throughout the AI lifecycle.

Furthermore, the findings of this review reinforce the imperative for the Computational Intelligence community not only to pursue more efficient algorithms but also to integrate ethical considerations from the outset — an approach aligned with ethics by design. This includes the development of robust fairness metrics for machine learning, the advancement of XAI techniques applicable across a variety of computational intelligence models, and the design of mechanisms to ensure meaningful human control over autonomous systems, particularly those based on reinforcement learning or evolutionary computation.

Overall, this review highlights both the maturity and fragmentation of the field: while conceptual clarity is improving, empirical applications of ethical frameworks in real-world industrial systems remain limited. Bridging this gap is a critical task for researchers, engineers, policymakers, and organizations striving to align technological innovation with human values and social responsibility.

VI. FINAL CONSIDERATIONS

The transition toward human-centric and ethically grounded AI systems represents one of the most critical challenges and opportunities of Industry 5.0. As this article has argued, the integration of artificial intelligence into industrial processes must go beyond efficiency and performance metrics to incorporate values such as fairness, transparency, and human well-being.

Through a critical review of recent literature and ethical frameworks, we have explored how responsible innovation can be guided by normative principles, technical transparency, and inclusive governance mechanisms. However, translating these ideals into practice remains an ongoing challenge. The persistence of ethical dilemmas — such as algorithmic bias, opacity, and accountability gaps — requires not only technical innovation but also institutional and cultural transformation.

The responsible deployment of AI in Industry 4.0 contexts will depend on robust regulation, participatory design processes, and a commitment to long-term societal goals. This includes fostering interdisciplinary collaboration and engaging stakeholders across the value chain — from engineers and operators to policymakers and civil society.

Ultimately, responsible AI is not a fixed destination but a continuous process of reflection, adaptation, and co-governance. By embedding ethical principles into the design,

deployment, and monitoring of AI systems, we can shape a future where technology serves not just productivity, but also human dignity and sustainability.

VII. Limitations and Future Research

Although this study has thoroughly explored the ethical, regulatory, and operational challenges of implementing artificial intelligence in Industry 4.0, several important limitations must be acknowledged. Firstly, the analysis focused predominantly on normative frameworks and theoretical proposals drawn from scientific and institutional literature. The absence of specific empirical case studies limits the assessment of the practical effectiveness of these models in real industrial environments, where organizational, cultural, and technological dynamics may create tensions or deviations between proposed principles and their day-to-day application.

In addition, the adopted approach emphasized a general and cross-sectoral perspective of Industry 4.0, which may fail to capture relevant nuances of specific sectors such as the food, energy, or logistics industries, where the impacts of AI present unique characteristics. The challenges of AI adoption in developing countries were also not explored in depth, despite institutional conditions, infrastructure, and technological maturity differing significantly from those of the core nations often cited in international guidelines.

Another limitation lies in the persistent difficulty of translating widely accepted ethical principles — such as fairness, transparency, and responsibility — into concrete technical and organizational requirements. This operationalization gap, as highlighted by several authors, underscores an urgent need for methodologies that integrate ethics from system design (ethics by design) to the continuous auditing of AI systems, especially in industrial contexts where automated decisions may directly affect safety, human labor, and the environment.

In this regard, future research should prioritize three complementary fronts. The first concerns empirical investigation into how industrial organizations are implementing — or failing to implement — ethical governance mechanisms, particularly regarding algorithmic accountability and human oversight. The second involves the development of technical tools and auditable metrics that enable real-time transparency and traceability of algorithmic decisions. Lastly, interdisciplinary and participatory studies that incorporate the perspectives of workers, affected communities, engineers, and policymakers are needed to build more inclusive and adaptable co-governance models.

By addressing these gaps, it will be possible to advance not only in the construction of more trustworthy systems aligned with ethical values but also in the consolidation of an organizational culture that understands AI not merely as an enabling technology, but as a strategic factor requiring continuous ethical commitment and distributed responsibility.

Moreover, the generalizability of our findings is constrained by publication bias and the prevalence of conceptual work. Future studies should pre-register protocols, report auditable fairness and explainability metrics in production settings, and release documentation artifacts (model/data cards) and decision logs for independent scrutiny [53], [54].

VIII. Acknowledgments

This work has been supported by the following Brazilian research agencies: FAPEMIG, CAPES, CNPq.

References

[1] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," Mar. 02, 2017, arXiv: arXiv:1702.08608. doi: 10.48550/arXiv.1702.08608.

[2] A. Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[3] E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 11, pp. 4793–4813, Nov. 2021, doi: 10.1109/TNNLS.2020.3027314.

[4] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The ethics of algorithms: Mapping the debate," Big Data & Society, vol. 3, no. 2, p. 2053951716679679, Dec. 2016, doi: 10.1177/2053951716679679.

[5] T. Papadakis, I. T. Christou, C. Ipektsidis, J. Soldatos, and A. Amicone, "Explainable and transparent artificial intelligence for public policymaking," Data & Policy, vol. 6, p. e10, Jan. 2024, doi: 10.1017/dap.2024.3.

[6] C. Högberg, "Stabilizing translucencies: Governing AI transparency by standardization," Big Data & Society, vol. 11, no. 1, p. 20539517241234298, Mar. 2024, doi: 10.1177/20539517241234298.

[7] Y. Zeng, E. Lu, and C. Huangfu, "Linking Artificial Intelligence principles," Dec. 12, 2018, arXiv: arXiv:1812.04814. doi: 10.48550/arXiv.1812.04814.

[8] H. Felzmann, E. Fosch-Villaronga, C. Lutz, and A. Tamò-Larrieux, "Towards Transparency by Design for Artificial Intelligence," Sci Eng Ethics, vol. 26, no. 6, pp. 3333–3361, Dec. 2020, doi: 10.1007/s11948-020-00276-4.

[9] M. J. Page et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," Mar. 2021, doi: 10.1136/bmj.n71.

[10] E. H. Grosse, Sgarbossa ,Fabio, Berlin ,Cecilia, and W. P. and Neumann, "Human-centric production and logistics system design and management: Transitioning from Industry 4.0 to Industry 5.0," International Journal of Production Research, vol. 61, no. 22, pp. 7749–7759, Nov. 2023, doi: 10.1080/00207543.2023.2246783.

[11] M. Tahaei, M. Constantinides, D. Quercia, and M. Muller, "A systematic literature review of human-centered, ethical, and responsible AI," Jun. 26, 2023, arXiv: arXiv:2302.05284. doi: 10.48550/arXiv.2302.05284.

[12] A. A. Khan et al., "Ethics of AI: A systematic literature review of principles and challenges," in The International Conference on Evaluation and Assessment in Software Engineering 2022, Gothenburg Sweden: ACM, Jun. 2022, pp. 383–392. doi: 10.1145/3530019.3531329.

[13] D. K. Gao, A. Haverly, S. Mittal, J. Wu, and J. Chen, "AI Ethics: A bibliometric analysis, critical issues, and key gaps," International Journal of Business Analytics, vol. 11, no. 1, pp. 1–19, Feb. 2024, doi: 10.4018/IJBAN.338367.

[14] L. Bibby and B. Dehe, "Defining and assessing industry 4.0 maturity levels – case of the defence sector," Production Planning & Control, Sep. 2018, doi: 10.1080/09537287.2018.1503355.

[15] R. S. Peres, X. Jia, J. Lee, K. Sun, A. W. Colombo, and J. Barata, "Industrial Artificial Intelligence in Industry 4.0 - Systematic review, challenges and outlook," in IEEE Access, vol. 8, pp. 220121–220139, 2020, doi: 10.1109/ACCESS.2020.3042874.

[16] Q. Zhu and J. Luo, "Toward artificial empathy for Human-Centered design: A framework," May 13, 2023, arXiv: arXiv:2303.10583. doi: 10.48550/arXiv.2303.10583.

[17] J. Morley, C. Machado, C. Burr, J. Cowls, M. Taddeo, and L. Floridi, "The debate on the ethics of AI in health care: A reconstruction and critical review," Nov. 13, 2019, Social Science Research Network, Rochester, NY: 3486518. doi: 10.2139/ssrn.3486518.

[18] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," Aug. 28, 2017, arXiv: arXiv:1708.08296. doi: 10.48550/arXiv.1708.08296.

[19] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," ACM Comput. Surv., vol. 51, no. 5, p. 93:1-93:42, Aug. 2018, doi: 10.1145/3236009.

[20] B. Shneiderman, "Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems," ACM Trans. Interact. Intell. Syst., vol. 10, no. 4, pp. 1–31, Dec. 2020, doi: 10.1145/3419764.

[21] R. Binns, "Fairness in machine learning: Lessons from political philosophy," in Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR, Jan. 2018, pp. 149–159. Accessed: May 01, 2025. [Online]. Available: https://proceedings.mlr.press/v81/binns18a.html.

[22] B. Mittelstadt, "Principles alone cannot guarantee ethical AI," Nat Mach Intell, vol. 1, no. 11, pp. 501–507, Nov. 2019, doi: 10.1038/s42256-019-0114-4.

[23] V. Eubanks, "Automating inequality: How high-tech tools profile, police, and punish the poor," Macmillan Publishers, Jan. 2018, ISBN: 9781250074317.

[24] C. D'Ignazio and L. F. Klein, "Data feminism". in Strong ideas series. Cambridge, Massachusetts: The MIT Press, 2020, ISBN: 9780262044004.

[25] A. Tóth, L. Nagy, R. Kennedy, B. Bohuš, J. Abonyi, and T. Ruppert, "The human-centric Industry 5.0 collaboration architecture," MethodsX, vol. 11, p. 102260, Dec. 2023, doi: 10.1016/j.mex.2023.102260.

[26] D. Long, "The new renaissance: can businesses build human-centric tech?," The Australian, Dec. 01, 2024. Accessed: May 02, 2025. [Online]. Available: https://www.theaustralian.com.au/business/growth-agenda/the-new-renaissance-can-businesses-build-humancentric-tech/news-story/6515c543d3ac9f622f497dd4f71e2f0b

[27] M. Breque, L. De Nul, A. Petridis, "Industry 5.0: Towards a sustainable, human-centric and resilient European industry," Luxembourg, LU: European Commission, Directorate-General for Research and Innovation, 2021. doi: 10.2777/308407.

[28] X. Xu, Y. Lu, B. Vogel-Heuser, and L. Wang, "Industry 4.0 and Industry 5.0 - Inception, conception and perception," Journal of Manufacturing Systems, vol. 61, pp. 530–535, Oct. 2021, doi: 10.1016/j.jmsy.2021.10.006.

[29] A. Adel, "Future of industry 5.0 in society: human-centric solutions, challenges and prospective research areas," Journal of Cloud Computing, vol. 11, no. 1, p. 40, Sep. 2022, doi: 10.1186/s13677-022-00314-5.

[30] AI HLEG, "Ethics guidelines for trustworthy AI," European Commission, Apr. 2019. Accessed: May 01, 2025. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[31] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices," Sci Eng Ethics, vol. 26, no. 4, pp. 2141–2168, Aug. 2020, doi: 10.1007/s11948-019-00165-5.

[32] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI," Jan. 15, 2020, Social Science Research Network, Rochester, NY: 3518482. doi: 10.2139/ssrn.3518482.

[33] I. Ahmed, G. Jeon, and F. Piccialli, "From Artificial Intelligence to Explainable Artificial Intelligence in Industry 4.0: A Survey on What, How, and Where," IEEE Transactions on Industrial Informatics, vol. 18, no. 8, pp. 5031–5042, Aug. 2022, doi: 10.1109/TII.2022.3146552.

[34] A. K. Badhan, R. Gill, "Explainable machine learning model for Industrial 4.0," in Machine Learning for Sustainable Manufacturing in Industry 4.0, CRC Press, 2023.

[35] K. Nikiforidis et al., "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, vol. 11, no. 1, pp. 135–148, Feb. 2025, doi: 10.1016/j.icte.2024.12.001.

[36] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," Aug. 09, 2016, arXiv: arXiv:1602.04938. doi: 10.48550/arXiv.1602.04938.

[37] J. M. Durán and K. R. Jongsma, "Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI," Journal of Medical Ethics, vol. 47, no. 5, pp. 329–335, May 2021, doi: 10.1136/medethics-2020-106820.

[38] J. Whittlestone, R. Nyrup, A. Alexandrova, and S. Cave, "The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions," in Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, in AIES '19. New York, NY, USA: Association for Computing Machinery, Jan. 2019, pp. 195–200. doi: 10.1145/3306618.3314289.

[39] European Commission, "Proposal for a Regulation laying down harmonised rules on artificial intelligence," European Commission. Accessed: May 02, 2025. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

[40] M. Veale and F. Z. Borgesius, "Demystifying the Draft EU Artificial Intelligence Act," Computer Law Review International, vol. 22, no. 4, pp. 97–112, Aug. 2021, doi: 10.9785/cri-2021-220402.

[41] A. F. T. Winfield and M. Jirotka, "Ethical governance is essential to building trust in robotics and artificial intelligence systems," Philos Trans A Math Phys Eng Sci, vol. 376, no. 2133, p. 20180085, Oct. 2018, doi: 10.1098/rsta.2018.0085.

[42] M. Ashok, R. Madan, A. Joha, and U. Sivarajah, "Ethical framework for Artificial Intelligence and Digital technologies," International Journal of Information Management, vol. 62, p. 102433, Feb. 2022, doi: 10.1016/j.ijinfomgt.2021.102433.

[43] V. Dignum, "Taking Responsibility," in Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way, V. Dignum, Ed., Cham: Springer International Publishing, 2019, pp. 47–69. doi: 10.1007/978-3-030-30371-6_4.

[44] A. Z. Coelho et al., "Governança da inteligência artificial em organizações: framework para comitês de ética em IA: versão 1.0," CEPI FGV Direito SP, May 2023. Accessed: May 02, 2025. [Online]. Available: https://hdl.handle.net/10438/33736

[45] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," ACM Comput. Surv., vol. 54, no. 6, p. 115:1-115:35, Jul. 2021, doi: 10.1145/3457607.

[46] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," Science, vol. 366, no. 6464, pp. 447–453, Oct. 2019, doi: 10.1126/science.aax2342.

[47] S. Barocas, M. Hardt, and A. Narayanan, "Fairness and machine learning," 2023. Accessed: Jan 08, 2025. [Online]. Available: https://fairmlbook.org/

[48] K. Holstein, J. Wortman Vaughan, H. Daumé, M. Dudik, and H. Wallach, "Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?," in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, in CHI '19. New York, NY, USA: Association for Computing Machinery, Maio 2019, pp. 1–16. doi: 10.1145/3290605.3300830.

[49] R. K. E. Bellamy et al., "AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias," Oct. 03, 2018, arXiv: arXiv:1810.01943. doi: 10.48550/arXiv.1810.01943.

[50] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and Abstraction in Sociotechnical Systems," in Proceedings of the Conference on Fairness, Accountability, and Transparency, in FAT* '19. New York, NY, USA: Association for Computing Machinery, Jan. 2019, pp. 59–68. doi: 10.1145/3287560.3287598.

[51] Q. N. Hong et al., "Mixed Methods Appraisal Tool (MMAT) version 2018: user guide," 2018. [Online]. Available: http://mixedmethodsappraisaltoolpublic.pbworks.com

[52] Q. N. Hong, V. Fàbregues, and P. Bartlett, "The Mixed Methods Appraisal Tool (MMAT) version 2022," 2022. [Online]. Available: http://mixedmethodsappraisaltoolpublic.pbworks.com

[53] M. Mitchell et al., "Model Cards for Model Reporting," in *FAT**, 2019, pp. 220–229. doi: 10.1145/3287560.3287596.

[54] T. Gebru et al., "Datasheets for Datasets," *Communications of the ACM*, 64(12), pp. 86–92, 2021. doi: 10.1145/3458723.

[55] M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," in *NeurIPS*, 2016, pp. 3315–3323.

[56] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR," *Harvard Journal of Law & Technology*, 31(2), 2018.

[57] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual Fairness," in *NeurIPS*, 2017, pp. 4066–4076.

[58] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *NeurIPS*, 2017, pp. 4765–4774.

[59] A. N. Angelopoulos and S. Bates, "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification," *arXiv:2107.07511*, 2021.

[60] M. Abadi et al., "Deep Learning with Differential Privacy," in *ACM CCS*, 2016, pp. 308–318. doi: 10.1145/2976749.2978318.

[61] ISO/IEC 42001:2023, "Artificial intelligence — Management system," International Organization for Standardization, 2023.

[62] Brasil, "Lei nº 13.709, de 14 de agosto de 2018 (Lei Geral de Proteção de Dados Pessoais)," 2018. [Online]. Available: https://www.planalto.gov.br/