ETHICAL DILEMMAS AND RISKS IN THE APPLICATION OF ARTIFICIAL INTELLIGENCE FOR INDUSTRIAL SAFETY: A SYSTEMATIC REVIEW.

Blanca Janeth Guaman Morocho[1], Andrés Alejandro Galvis Correa[2]

[1]Universidad Nacional de Chimborazo
[2]Universidad de las Fuerzas Armadas ESPE

janethgm1998@gmail.com[1]
aagalvis@espe.edu.ec[2]

Abstract
The incorporation of artificial intelligence (AI) in industrial safety in the manufacturing sector represents an emerging field with high potential for the proactive prevention of occupational risks, but it also entails ethical dilemmas and significant socio-technical risks. Through a systematic review of the literature (2015–2025) and a bibliometric analysis of 467 articles, this study identifies the main challenges related to worker privacy, fairness and algorithmic biases, autonomy and human control, as well as transparency and accountability. In addition, technical risks such as classification errors, model degradation (drift), cybersecurity vulnerabilities and reliance on edge/cloud architectures are documented. The findings reveal that Explainable AI (XAI), worker-centered socio-technical design, and the implementation of robust regulatory frameworks (AI Act, PLD, AILD) are key mitigation strategies. The study concludes that the integration of AI into industrial safety can optimize accident prevention as long as a comprehensive approach is implemented that articulates technical, organizational, and legal measures to ensure a safe, ethical, and equitablework environment.

Keywords: Artificial intelligence; industrial safety; ethics; algorithmic transparency; explainability; legal liability; psychosocial impact; manufacturing industry.

1. Introduction

1.1. Context of the Problem

The integration of artificial intelligence (AI) in industrial safety represents a paradigm shift, promising to move from reactive accident management to proactive risk prevention (Howard, 2019). The potential for AI systems to perform predictive analytics, monitor high-risk environments in real-time, and optimize security protocols is immense. However, this utilitarian promise is inextricably linked to a complex web of ethical dilemmas, legal challenges, and socio-technical risks. The fundamental tension between AI's optimising potential and the dangers inherent in its deployment – from invasive surveillance and perpetuation of bias to decision-making opacity and dilution of human responsibility – is at the heart of this research (European Agency for Safety and Health at Work [EU-OSHA], 2022).

1.2. Theoretical and Epistemological Framework

The implementation of artificial intelligence (AI) systems in the field of occupational safety is not a simple technological update; It is an intervention that profoundly recalibrates the power relations, responsibility structures, and core values that underpin the social contract in the workplace. This analysis deconstructs the ethical substrate of the problem, examining the moral principles that are challenged, the conceptual frameworks developed to govern AI, and the specific dilemmas that arise at the intersection of optimizing safety and individual rights.

1.2.1. Fundamental ethical principles under scrutiny

The deployment of AI in critical contexts has driven a remarkable international consensus around a core of ethical principles. Influential frameworks such as the European Commission, High-Level Expert Group on Artificial Intelligence (AI HLEG, 2019), the OECD Principles on AI (Organisation for Economic Co-operation and Development (OECD, 2024), and the UNESCO Recommendation on the Ethics of AI (2022) converge in a shared ethical vocabulary: respect for human autonomy, harm prevention, equity, transparency, and accountability (Floridi&Cowls, 2019).

However, underlying this apparent convergence lies a divergence in their philosophical foundations and strategic objectives. The EU framework is pragmatically geared towards consolidating the Digital Single Market under the banner of "European values" (AI HLEG, 2019, p. 4). The OECD's mission, faithful to its mission, has a predominantly socio-economic orientation, focused on inclusive growth and the explicit protection of "internationally recognized labour rights" (OECD, 2024). Finally, UNESCO's framework is anchored in a humanist universalism based on fundamental human rights (UNESCO, 2022). This divergence is not trivial; implies that the application of these principles in industrial safety may be controversial. A fatigue monitoring system could be justified from the perspective of market efficiency (EU), but be questioned from the perspective of labour rights (OECD) if it intensifies the pace of work, or from the perspective of human dignity (UNESCO) if it is considered excessively intrusive. This underlying tension is a critical source of practical ethical dilemmas.

1.2.2. The Locus of Responsibility: The Algorithmic Gap

Perhaps the most profound philosophical challenge posed by AI in high-risk contexts is the attribution of responsibility when an autonomous system fails and causes an accident. Andreas Matthias (2004) coined the term "responsibility gap" to describe the difficulty, or impossibility, of attributing blame to a specific human actor when a complex, learning-capable AI system causes harm in an unpredictable way. As direct human control over system operations diminishes, the traditional basis for the ascription of responsibilityfades (Nyholm, 2018).

Recent research delves into this concept, arguing that it is not a single problem, but "at least four interconnected problems" that encompass culpability, moral obligation, legal responsibility, and public responsibility (Santoni de Sio & Mecacci, 2021, p. 1057). This challenge materializes in an industrial scenario: a collaborative robotic arm, guided by reinforcement learning, performs an unforeseen movement and injures a worker. Neither the manufacturer, nor the company, nor the programmer can fully explain the algorithm's decision. In view of this, a key mitigating strategy is the implementation of "meaningful human control", which seeks to ensure that an informed human always maintains the ability to decide on the actions of the system (Santoni de Sio & van den Hoven, 2018). In a complementary way, the pragmatic approach of "algorithmic accountability" does not look for a moral culprit, but focuses on creating auditing, transparency, and governance mechanisms that ensure that there are avenues of redress for victims (Diakopoulos, 2016; Kroll et al., 2017).

1.2.3. The Productive Panopticon: Tension Between Security and Privacy

The continuous monitoring of workers using AI materializes one of the oldest tensions in political philosophy: the conflict between collective security and individual freedom. The utilitarian argument is that algorithmic surveillance can detect unsafe behaviour and prevent accidents, justifying a reduction in privacy for the sake of a greater good (EU-OSHA, 2022). However, a deontological perspective, anchored in rights, holds that privacy is an intrinsic value linked to human dignity and autonomy that should not be sacrificed (Ajunwa et al., 2017).

Empirical evidence confirms that intensive electronic surveillance is associated with higher stress and lower job satisfaction, eroding trust in the organization (Ravid et al., 2020). This conflict is magnified by the "power asymmetry" inherent in the employment relationship, a concern explicitly highlighted by the AI HLEG (2019, p. 10). European jurisprudence has addressed this tension, establishing that any workplace surveillance measure must be necessary and proportionate to the objective pursued, as determined in cases such as Bărbulescu v. Romania (2017) before the European Court of Human Rights. These philosophical dilemmas and rights tensions do not exist in a vacuum, but rather drive the creation of normative architectures designed to govern them.

1.3. Rationale and Research Question

The above epistemological analysis shows that the implementation of AI in industrial safety is at a crossroads of philosophical, ethical and fundamental rights debates. While conceptual frameworks and theoretical principles for ethical AI are increasingly defined internationally, there is a significant gap in understanding how these principles translate into practice within high-risk applied contexts such as the manufacturing sector. The theoretical literature warns about the dangers, but it is imperative to synthesize the available empirical and applied evidence to understand how these challenges manifest themselves and, crucially, what mitigation strategies are being proposed and evaluated. Therefore, a systematic literature review (SLR) is warranted to rigorously and transparently map existing evidence, identify gaps in research, and build an empirical foundation that can guide responsible implementation. Consequently, this research seeks to answer the following central question:

What are the ethical dilemmas and risks associated with the implementation of artificial intelligence in the industrial safety of companies in the manufacturing sector, and how can they be mitigated to ensure a safe and equitable work environment?

1.4. Hypothesis

Based on the theoretical tension identified between the optimizing potential of AI and its inherent risks, this study aims to validate the following research hypothesis:

The implementation of artificial intelligence in industrial safety within companies in the manufacturing sector can optimize the prevention of occupational risks, as long as ethical measures that protect worker privacy, strategies that mitigate adverse psychosocial impact, and governance frameworks that clearly define legal liability in the event of system failures are implemented concurrently and verifiablely.

2. Methodology

The present research has been executed using a three-phase protocol, designed to ensure both comprehensive coverage of the field and in-depth analysis of the evidence. This approach, called the "Triple Lens Strategy", combines a bibliometric analysis with a systematic literature review (SLR), reported in accordance with the PRISMA 2020 guidelines (Page et al., 2021a).

Phase 1: Corpus Search and Refinement (SALSA: Search)

The objective of this phase was to construct a corpus of highly relevant literature based on a systematic search and subsequent methodological refinement.

2.1. Sources and Search Strategy

A systematic search was conducted in the Web of Science (WoS) Core Collection database, covering the period from January 2015 to August 2025. A high-precision search equation was used that combines the key concepts of the research using Boolean and proximityoperators (NEAR/5).

("industrial safety" OR "occupational safety" OR "occupational health" OR "workplace safety" OR "risk prevention" OR "accident prevention" OR manufactur*) AND (("artificial intelligence" OR AI OR "machine learning" OR "deep learning" OR "algorithmic system*" OR "predictive analytics" OR "computer vision") NEAR/5 (ethic* OR privacy OR surveillance OR bias OR fairness OR transparen* OR explainab* OR XAI OR accountab* OR liability OR psychosocial OR cybersecurity))

The execution of this search yielded a total of 510 records.

2.2. Refinement of the Corpus by Thematic Categories

To increase the accuracy of the corpus, a filter based on the "Web of science Categories" was applied. Thematic categories considered irrelevant to the research question (e.g., biomedical sciences, earth sciences, pure physics) were systematically excluded. This refinement process, documented to ensure transparency, resulted in a refined corpus of 467 articles, which constituted the starting point fortheanalysisphases.

Phase 2: Scientific Mapping and Study Selection (SALSA: Appraisal)

Bibliometric analysis was applied to the refined corpus of 467 articles, followed by the content-based study selection process.

2.3. Bibliometric Analysis of the Refined Corpus

Landscape Analysis with bibliometrix: The bibliometrix package in R (Aria & Cuccurullo, 2017) was used to perform a performance analysis (identifying key authors, countries, and institutions) and a conceptual structure analysis (mapping thematic clusters based on theco-occurrence of keywords).

Analysis of Intellectual Structure with tosr: The "Tree of Science" (ToS) methodology (Robledo et al., 2024) was implemented to classify the 467 articles into Roots (foundational), Trunk (structural), Branches (sub-specializations) and Leaves (avant-garde), and thus analyze the intellectual influence structureofthefield.

2.4. Selection Criteria for Qualitative Synthesis

The following inclusion and exclusion criteria were defined for study selection:

Inclusion Criteria:

- Type of Document: Articles from scientificjournals and Reviews.
- Language: Documents written in English or Spanish.
- Thematic Content: The study must explicitly address the application or discussion of an AI technology in the context of industrial safety and analyze at least one ethical, legal, or sociotechnical dilemma of interest.

Exclusion Criteria:

- Purely technical studies without discussion of their implications.
- Opinion articles, editorials, book summaries or grey literature.
- Studies in a non-industrial sector without clear transferability.
- Studies not available in full text.

2.5. Study Selection Process

Two independent review authors conducted a two-stage screening of the 420 articles in the refined corpus:

1. Screening of Titles and Abstracts: The selection criteria were applied to the titles and abstracts for a first exclusion.
2. Full-Text Review: Articles that passed the first phase were evaluated in their full-text version for final selection.

Disagreements in both phases were resolved by consensus, with the intervention of a third reviewer if necessary.

Phase 3: Data Extraction and Synthesis (SALSA: Synthesis & Analysis)

This final phase focuses on the content analysis of the corpus of studies finally included.

2.6. Data Extraction and Quality Assessment

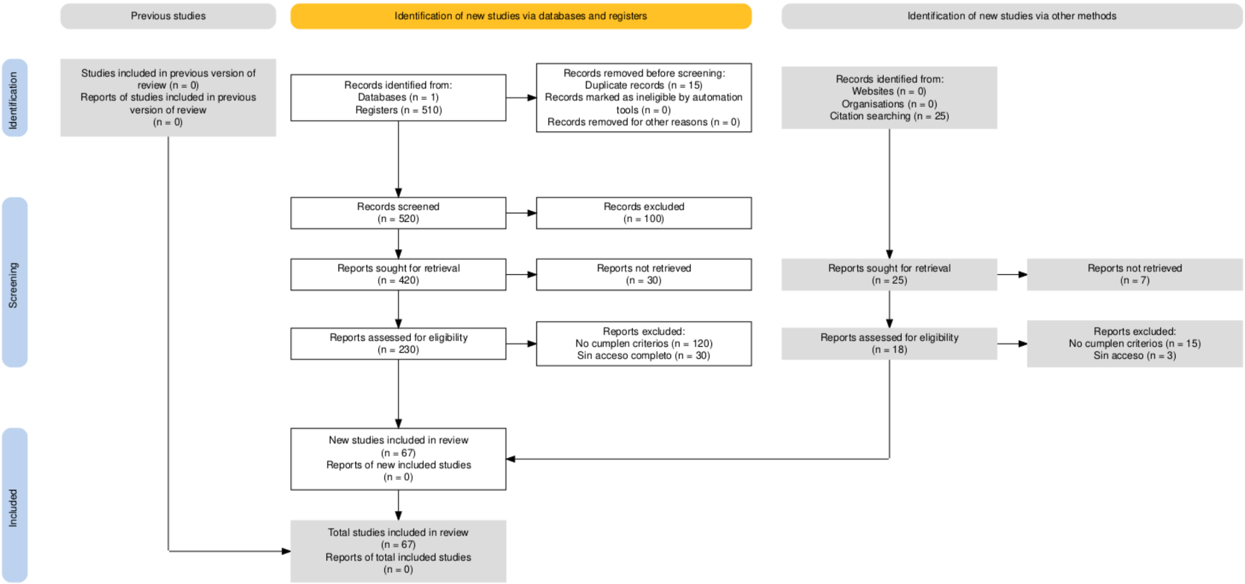
For each study in the final corpus, two review authors extracted the information using a standardised data sheet. A critical evaluation of the methodological rigor of each study was carried out, adapting the guidelines for literature reviews in heterogeneous fields (Snyder, 2019), in order to contextualize the robustness of theirfindings.

2.7. Synthesis Method

Given the heterogeneity of the studies, a thematic synthesis approach was employed. The extracted data were coded and grouped into analytical themes that directly align with the clauses of the research hypothesis: optimization, privacy, legal responsibility, and psychosocial impact.

3. Results

3.1. Selection of Studies.


Previous studies
Identification of new studies via databases and registers
Identification of new studies via other methods
Identification
Screening
Included
Studies included in previous version of review (n = 0)
Reports of studies included in previous version of review (n = 0)
Records identified from:
Databases (n = 1)
Registers (n = 510)
Records removed before screening:
Duplicate records (n = 15)
Records marked as ineligible by automation tools (n = 0)
Records removed for other reasons (n = 0)
Records identified from:
Websites (n = 0)
Organisations (n = 0)
Citation searching (n = 25)
Records screened (n = 520)
Records excluded (n = 100)
Reports sought for retrieval (n = 420)
Reports not retrieved (n = 30)
Reports sought for retrieval (n = 25)
Reports not retrieved (n = 7)
Reports assessed for eligibility (n = 230)
Reports excluded:
No cumplen criterios (n = 120)
Sin acceso completo (n = 30)
Reports assessed for eligibility (n = 18)
Reports excluded:
No cumplen criterios (n = 15)
Sin acceso (n = 3)
New studies included in review (n = 67)
Reports of new included studies (n = 0)
Total studies included in review (n = 67)
Reports of total included studies (n = 0)

Graph 1. PRISMA Protocol 2020

The bibliometric analysis of the collection from Web of Science, composed of 467 documents distributed in 278 sources, shows that research on the ethical dilemmas and risks associated with the implementation of artificial intelligence in industrial safety in the manufacturing sector constitutes an emerging field, characterized by a high annual growth rate (61.08%) and an extremely low average age of publications (1.48 years). which confirms its novelty and dynamism. The semantic richness derived from 809 Keywords Plus and 1,673 author keywords reflects a significant thematic granularity, supported by a solid theoretical foundation (26,943 references, 57.69 per document) that connects with consolidated disciplines such as ethics, engineering, occupational safety and AI technologies. The production, dominated by original articles (413) and complemented by reviews (54), shows patterns of intensive collaboration (4.43 co-authors per document) with a low proportion of individual authorship (7.5%), which underscores the need for multidisciplinary approaches. It highlights, however, the apparent absence of international co-authorships, a critical finding that, if confirmed, suggests an endogenous development bias that could limit the diversity of regulatory and cultural perspectives, and opens a strategic opportunity to foster transnational networks that strengthen the capacity of this field to generate ethical, safe and globally applicable solutions at the intersection between artificial intelligence and industrial safety.

The Three-Field Plot (CR–AU–ID) clearly exposes the cognitive architecture of the field, showing how seminal references in deep learning and transformers (Vaswani et al., 2017; Ke et al., 2017), together with foundational works on interpretability and explainable AI (Lundberg et al., 2017, 2020; Guidotti et al., 2019; Arrieta et al., 2020), constitute the epistemic basis from which central authors such as Mladenic, Fortuna, Kenda, Majlatow, and Rozanec are nourished, who translate this knowledge into applications oriented to Industry 4.0. On the thematic level, the most frequent keywords (model, framework, prediction, industry 4.0, security, fault-diagnosis, challenges) demonstrate a double anchorage: on the one hand, the adoption of high-performance AI architectures for process optimization and diagnosis; on the other, the explicit incorporation of explainability as a mechanism for mitigating risks and ethical dilemmas. The strong presence of terms such as security and fault-diagnosis confirms that the corpus not only pursues technical innovations, but is also oriented to the resolution of critical problems in industrial safety, in coherence with the research question. In summary, the analysis of this graph shows a flow of knowledge that goes from methodological advances in AI to specific industrial applications, with XAI as the backbone to guarantee transparency, robustness and trust in contexts where automated decisions have direct consequences on safety and labor equity. (see graph 2)

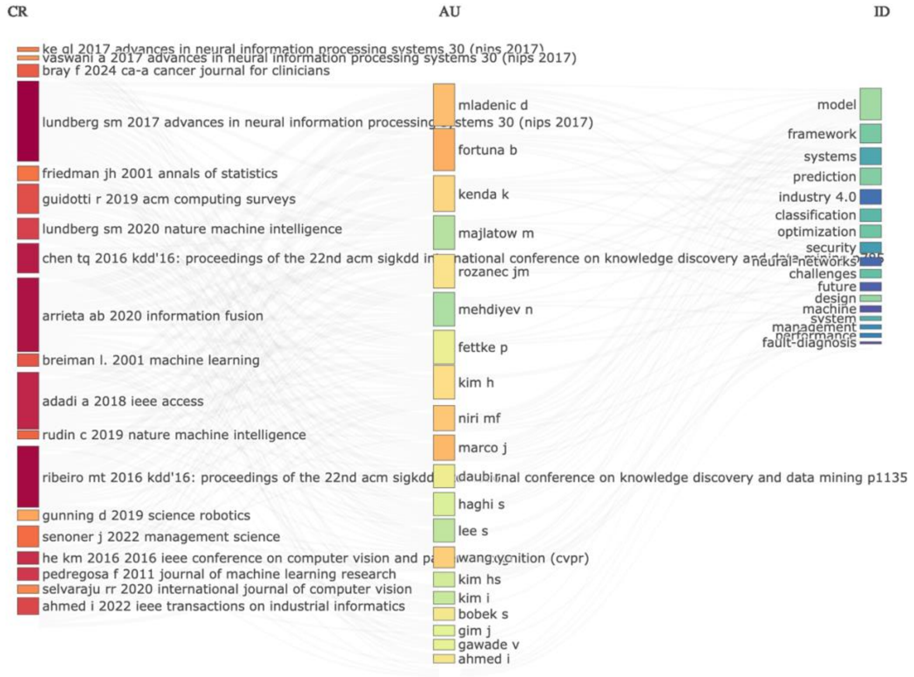
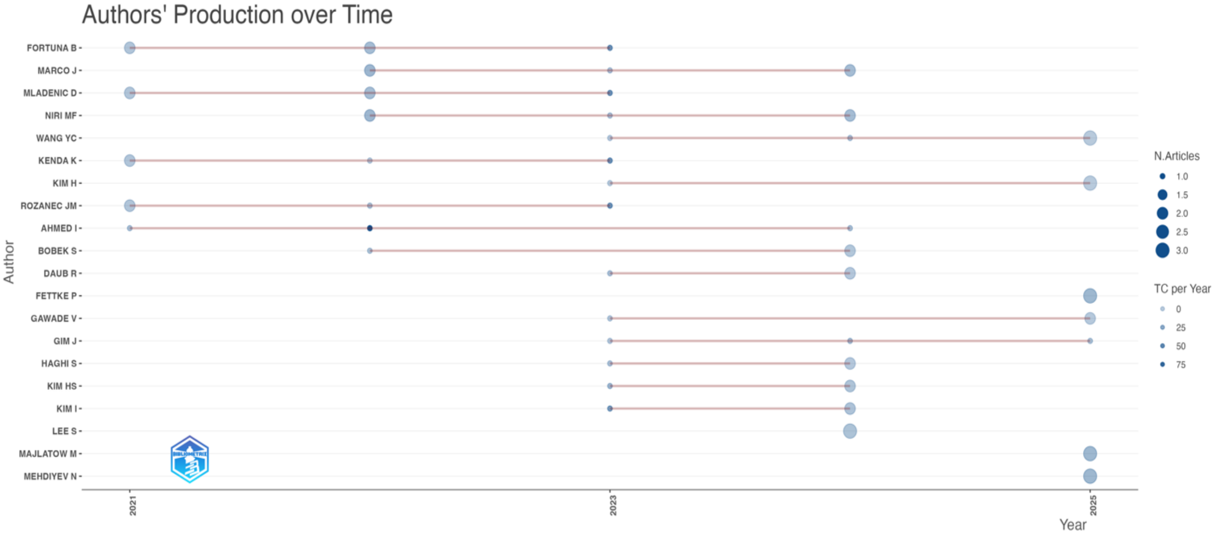

CR
AU
ID
ke ql 2017 advances in neural information processing systems 30 (nips 2017)
vaswani a 2017 advances in neural information processing systems 30 (nips 2017)
bray f 2024 ca-a cancer journal for clinicians
lundberg sm 2017 advances in neural information processing systems 30 (nips 2017)
friedman jh 2001 annals of statistics
guidotti r 2019 acm computing surveys
lundberg sm 2020 nature machine intelligence
chen tq 2016 kdd'16: proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining p785
arrieta ab 2020 information fusion
breiman l. 2001 machine learning
adadi a 2018 ieee access
rudin c 2019 nature machine intelligence
ribeiro mt 2016 kdd'16: proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining p1135
gunning d 2019 science robotics
senoner j 2022 management science
he km 2016 2016 ieee conference on computer vision and pattern recognition (cvpr)
pedregosa f 2011 journal of machine learning research
selvaraju rr 2020 international journal of computer vision
ahmed i 2022 ieee transactions on industrial informatics
mladenic d
fortuna b
kenda k
majlatow m
rozanec jm
mehdiyev n
fettke p
kim h
niri mf
marco j
daubi
haghi s
lee s
wang y
kim hs
kim i
bobek s
gim j
gawade v
ahmed i
model
framework
systems
prediction
industry 4.0
classification
optimization
security
neural-networks
challenges
future
design
machine
system
management
performance
fault-diagnosis

Figure 2 - Flow ofknowledgefrommethodologicaladvances in AI tospecific industrial applications

3.2. Production

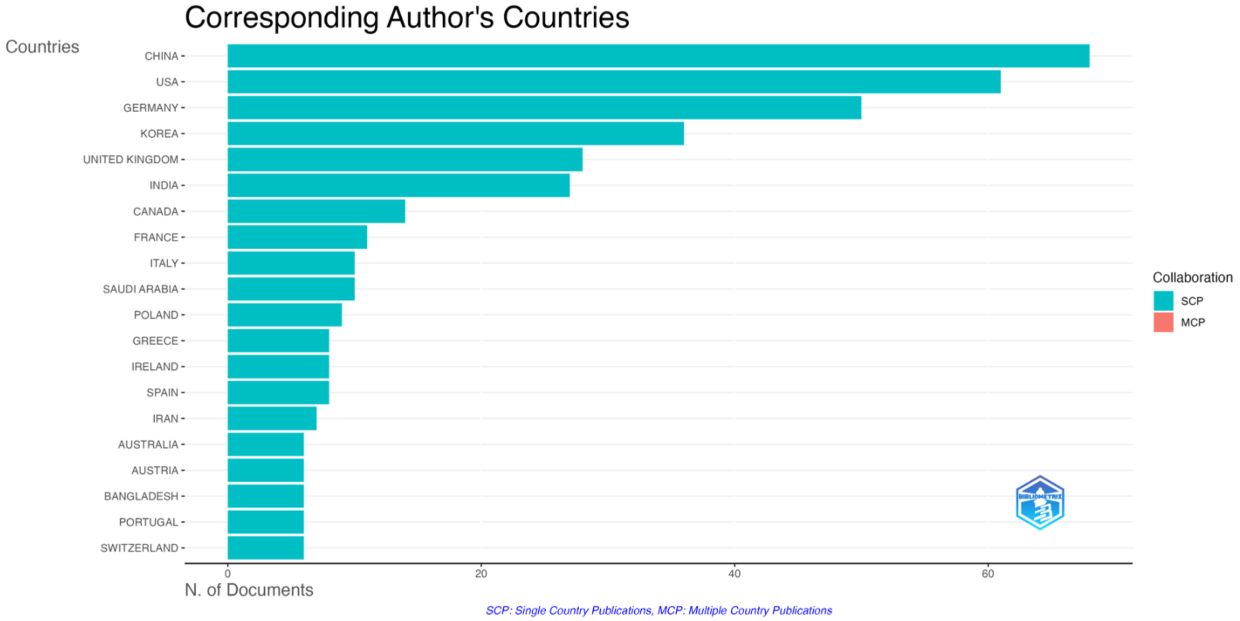
The analysis of the authors' production plot over time confirms that research on ethical dilemmas and risks of AI in manufacturing industrial safety is a frontier field, with a concentration of publications between 2021 and 2025 and an upward dynamism in both productivity and citation-normalized impact. Seminal contributions such as that of Ahmed I (2022), whose review in IEEE Transactions on Industrial Informatics constitutes a foundational framework in the transition from AI to explainable AI in Industry 4.0 (TCpY 93.8), and the group of Fortuna, Mladenic, Kenda and Rozanec (2023), with proposals for human-centered AI architectures (TCpY 42.7) aligned with Industry 5.0 principles. Alongside them, a spectrum of applied lines is observed that include multimodal occupational health surveillance (Kim I, 2023), the application of XAI to lithium battery manufacturing (Marco et al., 2022–2024), the prediction of defects in additive manufacturing (Kim HS, 2023), and the optimization of industrial processes through explainability (Gim J, Bobek S, Gawade V, 2023–2025). At the forefront of methodology, the projected works of Fettke, Majlatow and Mehdiyev (2025) introduce robust mechanisms for quantifying uncertainty and conforming prediction, aimed at reliability in safety-critical environments. Together, these trajectories show a consensus around Explainable AI (XAI) as a cross-cutting mitigation strategy, complemented by human-centric and algorithmic robustness approaches, which underscores the multidisciplinary and applied nature of a field that seeks to balance technological innovation, operational safety, and labor equity in thecontemporarymanufacturingindustry.



Graph 3- Authorproduction.

3.3. Authors

The analysis of the Corresponding Author's Country Collaboration Plot reveals that scientific production around the ethical dilemmas and risks of AI in manufacturing industrial safety is mainly concentrated in China (68 articles), the United States (61), Germany (50), Korea (36) and the United Kingdom (28), which confirms the role of these nations as research poles in the field. However, the most critical finding is the total absence of publications classified as multinational collaborations (CCM = 0% in the 20 most productive countries), which is anomalous in a frontier environment characterized by interdisciplinarity and the global nature of the challenges it addresses. This uniqueness may be due to a restrictive definition of the metric applied in Biblioshiny, to limitations of the extraction and normalization of affiliations in WoS, or to a real pattern of endogenous research development, conditioned by local regulations, availability of industrial data or predominantly national approaches. If the latter scenario is confirmed, it would imply a deficit of transnational exchange that could restrict cultural and regulatory diversity in the formulation of ethical and safe solutions. In any case, this finding requires rigorous methodological verification and triangulation with other international collaboration metrics, since the lack of cross-border co-authorship in such a globalized field constitutes an analytical gap with important implications for the consolidation of ethical standards and risk mitigation frameworks applicable onaninternationalscale.

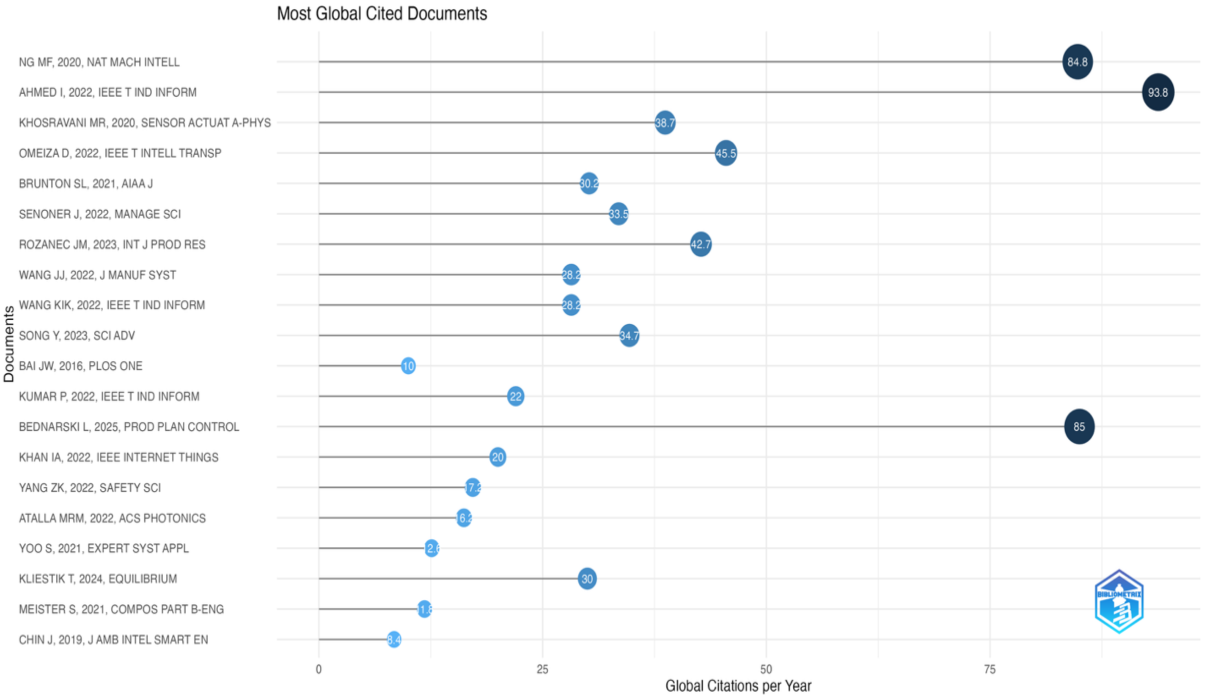


SCP: Single Country Publications, MCP: Multiple Country Publications

Graph 4 - Productionby country

3.4. Documents

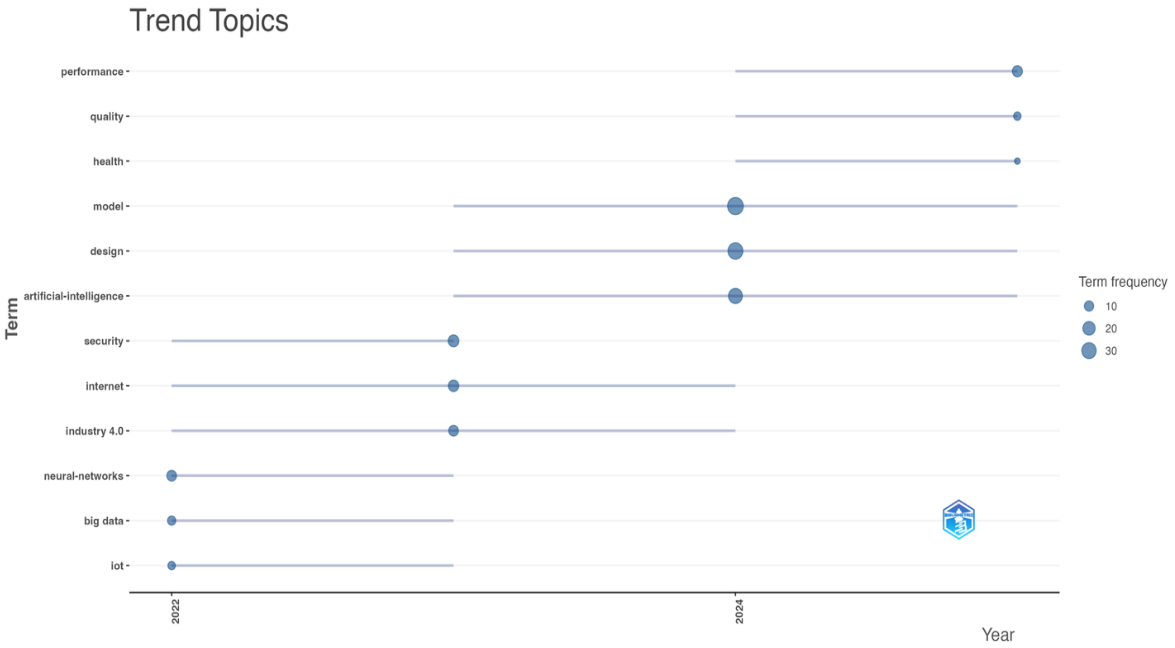
The Most Global Cited Documents graph shows that the intellectual structure of the field on ethical dilemmas and risks of AI in industrial safety is articulated around a small core of publications with very high global influence, led by Ahmed I (2022, IEEE T IND INFORM, GCpY 93.8) and Ng MF (2020, Nat Mach Intell, GCpY 84.8), which act as conceptual pillars around Explainable AI (XAI) and its application in industrial environments. In turn, works such as Omeiza D (2022, IEEE T Intell Transp, GCpY 45.5) and Rozanec JM (2023, Int J Prod Res, GCpY 42.7) consolidate the transfer of these approaches to smart manufacturing and transportation, while Senoner J (2022, Manage Sci) and Song Y (2023, Sci Adv) expand the debate towards organizational management and health surveillance, linking to the axes of psychosocial impact and legal responsibility. The presence of projected documents (e.g. Bednarski L, 2025, Prod Plan Control, GCpY 85) reinforces the emerging and prospective condition of the field, indicating that the future agenda will be marked by the robustness and integration of XAI in process planning and control. In summary, the graph confirms that intellectual influence is concentrated at the intersection between XAI, smart manufacturing and risk management, constituting the backbone on which the most recent applications and developments in thearea are structured.



Graph 5 - Productionbydocuments

3.5. Topics

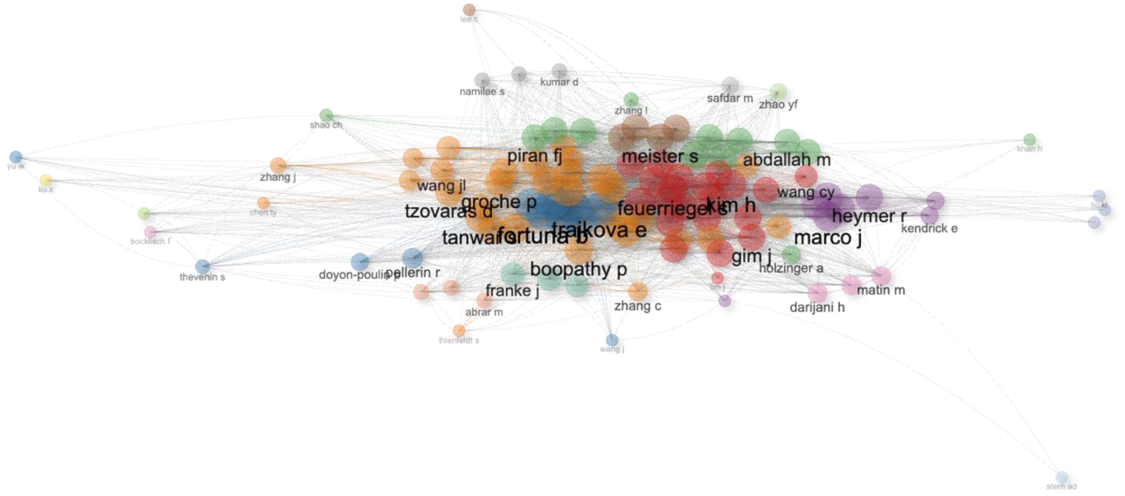
The Trend Topics chart shows that research on the ethical dilemmas and risks of AI in manufacturing industrial safety has transitioned, in a short period (2022–2024), from an initial emphasis on foundational technologies and the digital framework of Industry 4.0 (IoT, big data, neural-networks, security) to a more mature approach in the design and implementation of AI systems (model, design, artificial-intelligence), accompanied by metrics with a direct impact on occupational safety and well-being (health, quality, performance). This evolution reveals that the scientific community has gone from discussing the technological substrate to problematizing model configurations and their practical implications in terms of risks, transparency, and mitigation. Although terms such as ethics or fairness do not explicitly emerge among the most frequent, it is likely that they are subsumed in technical categories such as "design" and "model". Overall, the graph shows a rapidly consolidating field, where risk mitigation is conceived not only as a cybersecurity or connectivity problem, but as a matter of responsible design and evaluation of results, which connects directly to the need for safe and equitable worke nvironments.

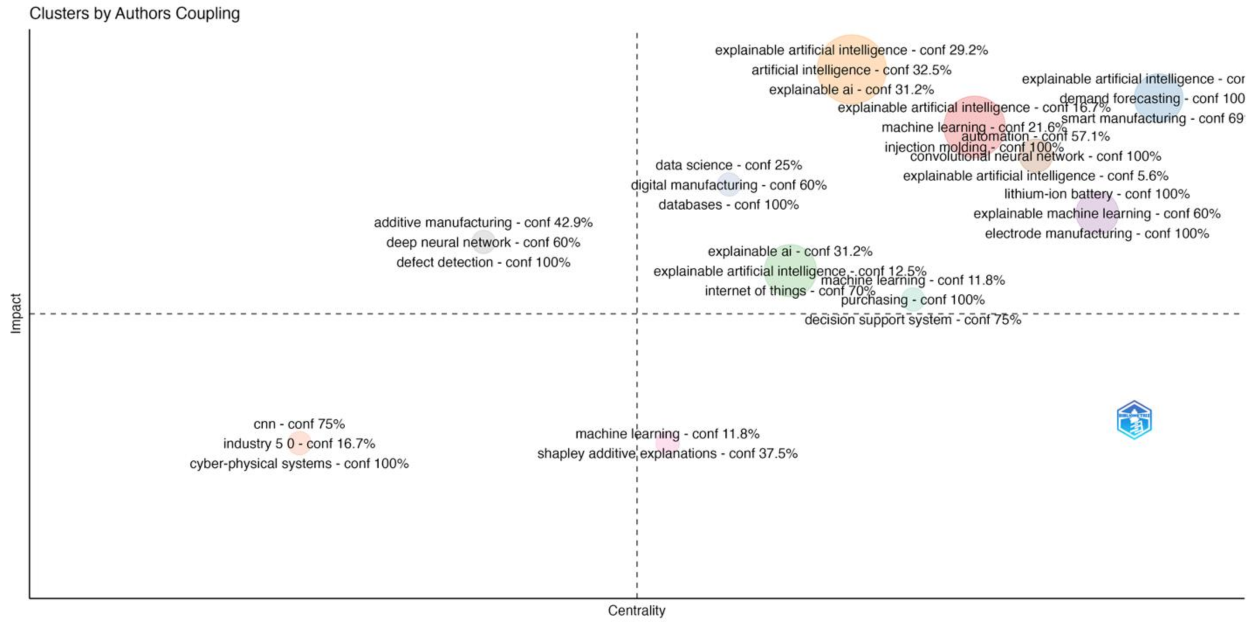


Graph 6 - Productionbytopic

3.6. Cluster

The joint analysis of the clusters by authors coupling graphs, the co-authorship network and the topic table confirms that the field of research on AI, ethics and manufacturing industrial safety is structured around a robust methodological core in XAI and machine learning, complemented by high-risk applied verticals such as additive manufacturing, defect detection, injection molding, and lithium batteries. Clusters located in the high centrality and impact quadrant (e.g., explainable AI, smart manufacturing, automation, demand forecasting) constitute the core of the Tree of Science, while specialized topics with strong impact but less integration (defect detection, electrode manufacturing) represent critical branches of application. The network of co-authorships reveals a dense and multipolar ecosystem, with authors such as Fortuna, Mladenic, Trájkova, Meister and Kim acting as central nodes and brokers interconnecting communities, which evidences a real collaboration that contrasts with the zero value of MCP reported on Biblioshiny and suggests reviewing the affiliation metric. Together, these results reinforce that the research agenda is migrating from base technologies towards explainable, traceable and safety-focused models, aligned with the synthesis categories of your protocol (optimization, privacy and security, legal responsibility and psychosocial impact), consolidating a rapidly maturing field with clear lines of research for risk mitigation and the guarantee of safe and equitableworkenvironments.

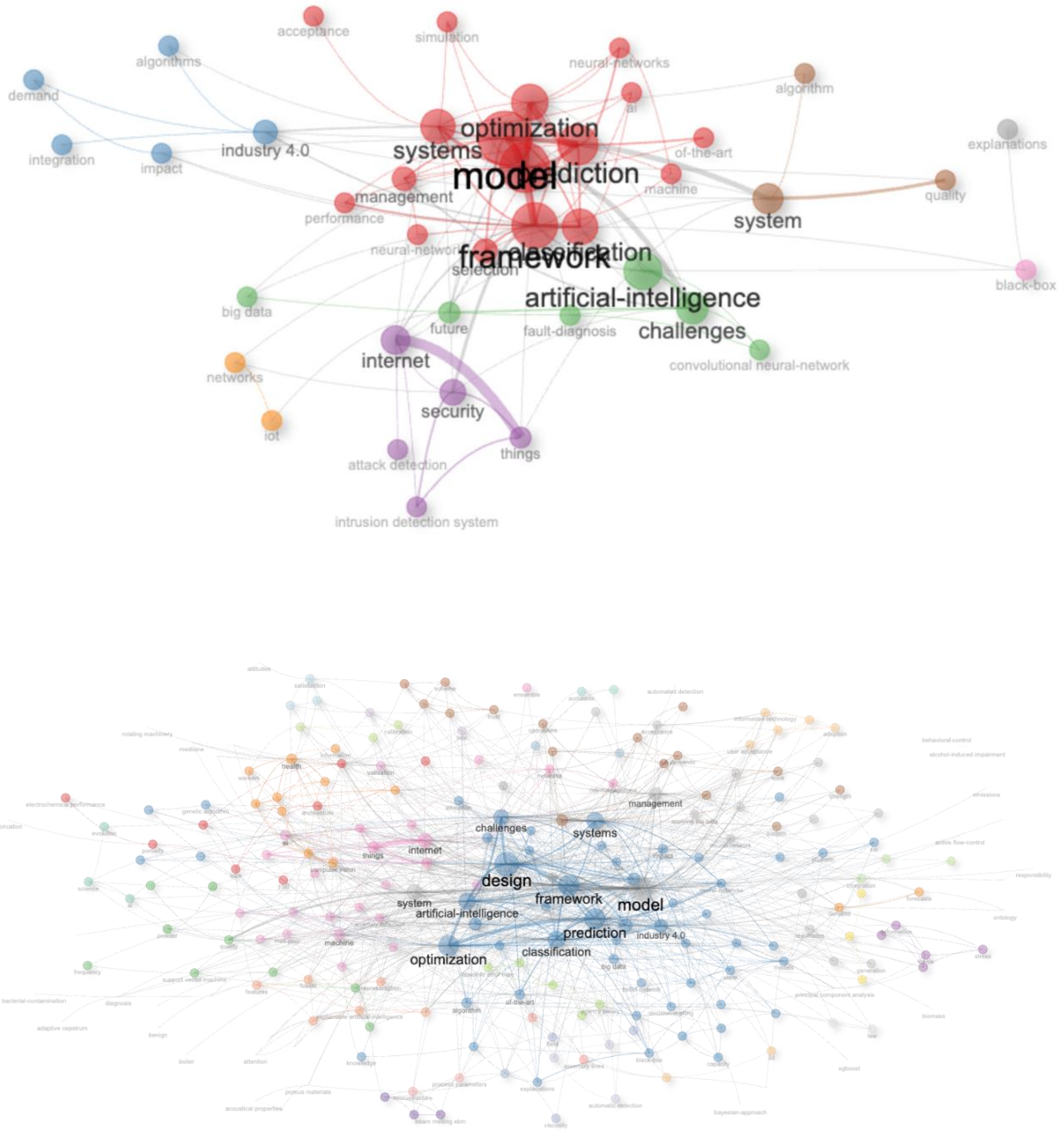




Graph 7 - Clusters

3.7. Conceptual Co-occurrence

The analysis of the Keywords Plus word co-occurrence network reveals that the literature on ethical dilemmas and risks of AI in manufacturing industrial safety is organized around several well-defined clusters: a technological one focused on the development and optimization of models (models, systems, frameworks, neural networks, optimization and performance), another conceptual and applied one that directly addresses the challenges of artificial intelligence and fault diagnosis (artificial-intelligence, challenges, fault-diagnosis), a third core oriented to cybersecurity and intrusion detection (security, IoT, attack detection, intrusion detection system), a contextual-industrial cluster that frames the discussion on Industry 4.0 and its impact, and a critical group on explainability and the "black-box" (explanations, black-box, quality), which is at the heart of ethical dilemmas. The structure of the network shows that research has evolved from the construction of technical architectures to the incorporation of transparency, trust and acceptance mechanisms, integrating both cybersecurity and explainability as risk mitigation strategies. In this sense, the map confirms that academic debates are not limited to the efficiency of algorithms, but are increasingly focused on the need for robust, auditable and accountable systems, which are essential to ensure safe and equitable work environments in the framework of the digital transformationofmanufacturing.



Graph 8 - Networks

3.8. Synthesis: ResearchQuestions and Hypotheses

1) Ethical dilemmas

Privacy and surveillance in the workplace.

Vision/IoT systems for security (detection of insecure postures, PPE, access, geolocation) involve a massive capture of personal and biometric data. The literature shows that it can be mitigated with privacy-by-design (e.g., face blurring, "pose-only pipelines") that preserve utility for security by reducing worker identification.

When analytics are moved to the edge or trained with federated learning (data does not leave the plant and only travels in gradients), the exposure of sensitive data is reduced, maintaining accuracy in industrial tasks.

In addition, differential privacy at the stage of training or release of synthesized data introduces calibrated noise that protects identities without completely destroying the usefulness of the model.

At the organizational level, Frameworks for Trustworthy AI (NIST AI RMF; European guidelines) call for proportionality, data minimization, and lifecycle governance.

Equity, bias and non-discrimination.

In manufacturing, bias can emerge from unrepresentative data (e.g., operators, shifts, lines, or underrepresented tasks). The review in Sensors emphasizes that "bias and accountability" are critical dimensions of industrial trustworthiness and that they need to be systematically audited, explained, and mitigated.

Autonomy, human control and "dehumanization".

The I5.0 papers point out that technical research overshadowed psychosocial factors: confidence, autonomy, and operator stress are underserved and, if ignored, undermine the safe and legitimateadoptionof AI.

Trust calibration warns that a "poorly calibrated" trust (excess or deficit) produces errors: over-confidence → omission of checks; sub-trust → rejection of the system. Explicit design of signals, explanations, and operational boundaries is required to align expectations.

Transparency and explainability.

For industrial safety, explainability is not a "luxury": it is a condition for audit, worker acceptance and legal expertise. I5.0/XAI revisions list interpretability as a pillar for human-centered and plant-assessable AI.

2) Technical and operational risks

False positives/negatives with an impact on security.

Task and posture classifiers applied to prevention can fail due to environmental conditions (lighting, occlusions, heterogeneous footwear/equipment, human variability), which translates into erroneous alarms or non-detection of dangerous behaviors and, therefore, residual risks. The investigations also report workers' concerns about cameras and intrusion, which affect compliance with measures and the legitimacy of the system.

Drift degradation and plantchanges.

In discrete environments and continuous processes, model performance degrades over time (new products, stations, PPE, layout, lighting). The manufacturing literature highlights the need to monitor models, recalibrate and version data/algorithms under risk management frameworks (NIST/ISO).

Cybersecurity of data models and pipelines.

Attack surfaces include cameras, gateways, OT/IT networks, and the ML pipeline itself. Defense in-depth controls and specific tests aligned with AI risk frameworks are recommended.

Edge/cloud integration and latencies.

For real-time security analysis, latencies or outages can leave "blind zones." Federated/on-premise architectures for predictive maintenance and monitoring (e.g., industrial wind farms, analog in manufacturing) illustrate how distributing compute reduces data exposure and communications bottlenecks.

3) Legal frameworks and responsibility

Complementarity between regulation and responsibility.

The European strategy articulates an "ecosystem of trust" where the AI Act establishes ex ante requirements (risk management, documentation, human supervision), and the civil liability directives (revision of the AML and AI Liability Directive) facilitate ex post compensation and adjust incentives.

The doctrinal analysis highlights that the burden of proof and technical opacity complicate access to compensation; hence the need to alleviate evidence, expand definitions of "product" to include software/services, and articulate defectiveness principles in line with security expectations.

Efficient risk allocation.

The proposal for "two-track" product liability suggests supplementing default liability with a truly objective regime for particularly dangerous products (e.g., autonomous systems), distributing costs to the "cheapest avoider" withoutstiflinginnovation.

Economic models of hybrid rules show that, when certain activities are verifiable (e.g., operator or manufacturer), rules with due levels of activity and care can achieve social optima; "Objective double liability" is also discussed with mechanisms to align investments in security.

The revised U.S. doctrine identifies the three classic triggers (manufacturing defect, design, and instruction/warning) and their limits when the system "works as designed," just the hard case in industrial AI.

4) Psychosocial impact and acceptance

Confidence, autonomy and stress.

The evidence from I5.0 shows that the psychosocial dimension has been underaddressed; Experimental, multi-method research is called for to measure confidence, motivation, stress, and designworker-centeredsystems.

The Trust Calibration Framework proposes to design signals, explanations, and usage limits that maintain trust "in the safe zone" (neither blind nor cynical).

Studies of "AI in physical tasks" report discomfort with cameras and perceptions of surveillance; if this is not managed, it impacts the safety climate and adherence.

5) Mitigation strategies (technical, organizational, and legal)

A. Data privacy and governance (by design)

- Anonymous and minimal vision: blur faces, use body keypoints, and edge processing before any transmission.
- Training without centralized data: federated learning with secure aggregation and client drift auditing.
- Differential privacy and data synthesis for industrial-grade datasets.

B. Technical reliability and safety MLOps

- Performance and drift monitoring, periodic recalibration, data/model versioning with NIST AI RMF/ISO controls.
- Specific safety tests (FPR/FNR on critical classes), adversarial tests and documentedoperationalkill-switch.

C. Transparency, Explainability and Auditing

- Actionable XAI (local rationality per security event, traceability of sensors to decision).
- Decision logs and accountability ontologies (e.g., outlining requirements, evidence, and roles) for expertise and compliance.

D. Socio-technical design and change management

- Reliable calibration (design of alerts, thresholds, human feedback and "two-handed" procedures for critical actions).
- Risk-adaptive autonomy levels (e.g., P3LS: from passive policies to active monitoring), with human-in-the-loop on high-impact tasks.
- Training programs, co-design with the operator, and digital worker bills of rights (clarity on what is beingmeasured and why).

E. Regulatory framework and contracts

- Implement AI Act requirements (risk management, conformity assessment, human oversight) and align policies/contracts with revised AML and strict liabilityproposalswhereapplicable.
- Internal evidence logging policies to facilitate evidentiary burdens and audits in incidents (defectiveness, causal link).

6) Summary Answer to the Central Question

1. Ethical dilemmas: privacy (capture/use of biometrics), biases and equitable treatment, loss of autonomy/stigmatization, opacity and diffuse responsibility. (Supported by privacy by design, XAI, and trust frameworks.)
2. Technical/operational risks: classification errors with safety consequences; drift and degradation; cyberattacks on the pipeline; Real-time integration/latency failures. (Evidence in industry reviews and field studies.)
3. Comprehensive mitigation:

or Technology: Anonymous Vision, FL, Differential Privacy, XAI + Security MLOps.
or Organization: reliable calibration, risk autonomy levels (P3LS), training and co-designwithoperators.
or Legal/governance: AI Act (ex ante) + AML/AILD (ex post), and—where applicable—complementary two-track strict liability regime forparticularlydangerousproducts.

This package reduces data exposure, improves robustness and verifiability, and clarifies the assignment of responsibilities, enabling more effective and equitable industrial safety in manufacturing plants.

The implementation of AI in industrial safety in the manufacturing sector offers benefits for risk prevention and process optimization, but it poses ethical dilemmas and risks that must be managed with a comprehensive approach.

1. Main ethical dilemmas:

- Privacy and surveillance: Bulk collection of biometric and behavioral data of workers.
- Fairness and bias: risk of discriminatory decisions due to incomplete or unrepresentative data.
- Autonomy and human control: possibility of dehumanization of work and loss of agency of the worker.
- Transparency and trust: opacity of algorithms that makes accountability difficult.

2. Technical and operational risks:

- Classification errors: false positives/negatives that impact real security.

- Model degradation (drift): loss of accuracy due to changes in theplantenvironment.
- Cybersecurity: vulnerability to attacks on sensors, networks and models.
- Technological dependence: overconfidence or rejection of intelligent systems.

3. Mitigation strategies:

- Techniques: computer vision with anonymization (blurring, keypoints), federated learning, differentialprivacy, XAI audits.
- Organizational: trust calibration, human-in-the-loop supervision, training programs, and co-designwithworkers.
- Legal and governance: implementation of the AI Act (risk prevention and management), revision of the Product Liability Directive and adoption of strict liability modelsforhigh-riskproducts.

4. Discussion

The results of this systematic review confirm that the implementation of artificial intelligence (AI) in industrial safety is an expanding, dynamic and multidisciplinary field. The bibliometric analysis shows an accelerated growth in scientific production, concentrated in the last five years, which reflects the growing importance of the topic within the framework of Industry 4.0 and the transition to Industry 5.0. However, it also shows significant limitations, such as the lack of international collaboration among researchers, which can restrict the diversity of regulatory and cultural approaches necessary for the creation of universal ethical frameworks.

Ethical and philosophical dimension

One of the central findings is that AI poses ethical dilemmas of great relevance in the work environment. Mass surveillance and the use of biometric data, while they can help prevent accidents, also generate tensions with the fundamental right to privacy and dignity of workers. The literature reviewed highlights the need to adopt theprincipleof*privacybydesign*, incorporatingmechanismssuch as anonymization, federated learning or differential privacy. These technical approaches are not only safety measures, but strategies that allow for the articulation of an ethical framework where the well-being of the worker is as much a priority as productive efficiency.

It is also observed that algorithmic biases constitute a real risk in the manufacturing sector. When the data used to train models does not adequately represent the diversity of operators, shifts or contexts, there is a risk of generating discriminatory decisions. This connects to broader philosophical debates about justice and equity, raising the need for systematic bias audits and shared responsibility frameworks between manufacturers, employers, and algorithm designers.

Technical and operational dimension

From a technical point of view, the study shows that the adoption of AI in industrial safety is not without limitations. The appearance of false positives and negatives in detection systems can have critical consequences: a false negative may fail to detect unsafe behavior, putting a worker's life at risk; while a false positive can generate constant alarms that, in the long run, reduce employees' trust and commitment to the system.

The phenomenon of *concept drift*, i.e. the loss of accuracy of models as working environments change, is another operational challenge that requires continuous updating and recalibration strategies. Added to this is the increasing vulnerability of systems to cyberattacks in connected industrial environments, where sensors, cameras, and OT/IT networks represent critical attack surfaces. In this sense, the adoption of robust security MLOps frameworks, including resilience testing, adversarial testing, and documented*kill-switch* mechanisms, becomesessential.

Legal and liability dimension

On the legal level, the results confirm the difficulty of attributing responsibility for accidents caused by autonomous systems. The concept of the "responsibility gap" remains an unsolved problem in practice, especially when algorithmic decisions cannot be fully explained. European regulatory proposals, such as the **AI Act** and the revisionofthe**ProductLiability Directive (AML),** are importantadvances in articulating a system that combines ex-ante requirements (risk management, documentation, human supervision) with ex-post compensation mechanisms. However, gaps still remain on how to apply these frameworks to rapidly evolving technologies.

The academic debate also suggests exploring hybrid models of strict liability, especially for high-risk autonomous systems. This approach seeks to redistribute costs to the actors with the greatest mitigation capacity (manufacturers or insurers) without stifling technological innovation, balancing development incentives and worker protection.

Psychosocial and organizational dimension

Another key finding is that psychosocial factors have been underestimated in the literature. The introduction of AI systems in the workplace not only affects physical safety, but also the perception of autonomy, the level of trust in the organization, and the emotional well-being of workers. Recent studies warn that intensive surveillance can increase stress, reduce job satisfaction, and erode trust in the company.

Calibrating trust in intelligent systems is an essential condition for safe adoption. Poorly calibrated trust—whether excessive or undertrusted—can lead to dangerous behaviors: for example, overconfidence could lead workers to ignore safety protocols, while excessive distrust can lead to rejection of technologies that are intended to protect them in the first place. Therefore, it is necessary to design systems with understandable explanations, clear limits of autonomy and active supervision procedures that reinforce trust withoutreplacing human agency.

Future projection

Finally, the results of this study suggest that the future research agenda should focus on three main fronts:

1. **Multidisciplinary integration**: combining technical, legal, ethical and psychosocial perspectives in the design of AI systems.
2. **International collaboration**: fostering transnational networks that allow for the construction of more universal ethical and regulatory standards.
3. **Applied experimentation**: moving towards empirical studies that measure not only the technical effectiveness of systems, but also their impact on safety culture, workplace well-being and social acceptance.

Overall, the discussion underlines that AI in industrial safety should not be conceived solely as a technological tool, but as a socio-technical transformation that requires governance, ethical design, and an explicit commitment to worker dignity.

5. Conclusions

The integration of artificial intelligence in industrial safety offers a double face: on the one hand, it makes it possible to detect risks early, optimise processes and reduce accidents; on the other, it raises ethical dilemmas, technical risks, and legal challenges that cannot be ignored.

This study concludes that the responsible adoption of AI in the manufacturing industry depends on three fundamental pillars:

1. **Technological**: use of techniques such as anonymous vision, federated learning, differential privacy, and algorithmicexplainability.
2. **Organizational**: strengthening of meaningful human control, trust calibration, training programs, and active participation of workers in the design of systems.
3. **Regulatory and legal**: alignment with international regulatory frameworks, adoption of strict liability models in high-risk systems, and clear audit and accountability mechanisms.

In short, AI can become a strategic ally of industrial safety only if it is implemented under a comprehensive, multidisciplinary approach focused on the dignity of the worker.

References

European Agency for Safety and Health at Work (EU-OSHA). (2022). Artificial intelligenceforworkermanagement and itsimplicationsforoccupational safety and health. Publications Office oftheEuropeanUnion.

Howard, J. (2019). Artificial intelligence: Implications for the future of work. American Journal of Industrial Medicine, 62(11), 917-926. https://doi.org/10.1002/ajim.23037.

Ajunwa, I., Crawford, K., & Schultz, J. (2017). Limitlessworkersurveillance. California LawReview, 105(3), 735–776.

Bărbulescu v. Romania [GC], no. 61496/08, ECHR 2017.

Diakopoulos, N. (2016). Accountability in algorithmicdecisionmaking. Communicationsofthe ACM, 59(2), 56-62. https://doi.org/10.1145/2844110

EuropeanCommission, High-Level Expert Groupon Artificial Intelligence (AI HLEG). (2019). Ethicsguidelinesfortrustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Floridi, L., &Cowls, J. (2019). A unifiedframeworkoffiveprinciplesfor AI in society. Harvard Data ScienceReview, 1(1). https://doi.org/10.1162/99608f92.8cd550d1.

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. University of Pennsylvania LawReview, 165, 633-705.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology, 6(3), 175-183. https://doi.org/10.1007/s10676-004-3422-1.

Nyholm, S. (2018). Attributingagencytoautonomoussystems: Reflectionsonthe moral and legal significanceof machine learning. Science and EngineeringEthics, 24(4), 1201-1219. https://doi.org/10.1007/s11948-017-9937-2.

OrganisationforEconomic Co-operation and Development (OECD). (2024). OECD AI Principles. https://oecd.ai/en/ai-principles

Ravid, D., Tomczak, A., Freeman, G., &Fruchter, N. (2020). PrivateEyes, TheySeeYourEveryMove: WorkplaceSurveillance and WorkerWell-Being. Academyof Management Proceedings, 2020(1).

Santoni de Sio, F., &Mecacci, G. (2021). Fourresponsibility gaps with artificial intelligence: Whytheymatter and howtoaddressthem. Philosophy&Technology, 34, 1057–1084. https://doi.org/10.1007/s13347-021-00450-x.

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control overautonomoussystems: A philosophicalaccount. Frontiers in Robotics and AI, 5, 15. https://doi.org/10.3389/frobt.2018.00015.

UNESCO. (2022). Recommendation on the Ethics of Artificial Intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000381137.