



Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/229564/>

Version: Published Version

Article:

Ahangar, M.N., Farhat, Z.A. and Sivanathan, A. (2025) AI trustworthiness in manufacturing: challenges, toolkits, and the path to Industry 5.0. *Sensors*, 25 (14). 4357. ISSN: 1424-8220

<https://doi.org/10.3390/s25144357>

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

<https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Review

AI Trustworthiness in Manufacturing: Challenges, Toolkits, and the Path to Industry 5.0

M. Nadeem Ahangar , Z. A. Farhat  and Aparajithan Sivanathan 

AMRC North West, University of Sheffield, Blackburn BB2 7HP, UK; z.farhat@amrc.co.uk (Z.A.F.); a.sivanathan@amrc.co.uk (A.S.)

* Correspondence: m.n.ahangar@amrc.co.uk

Abstract

The integration of Artificial Intelligence (AI) into manufacturing is transforming the industry by advancing predictive maintenance, quality control, and supply chain optimisation, while also driving the shift from Industry 4.0 towards a more human-centric and sustainable vision. This emerging paradigm, known as Industry 5.0, emphasises resilience, ethical innovation, and the symbiosis between humans and intelligent systems, with AI playing a central enabling role. However, challenges such as the “black box” nature of AI models, data biases, ethical concerns, and the lack of robust frameworks for trustworthiness hinder its widespread adoption. This paper provides a comprehensive survey of AI trustworthiness in the manufacturing industry, examining the evolution of industrial paradigms, identifying key barriers to AI adoption, and examining principles such as transparency, fairness, robustness, and accountability. It offers a detailed summary of existing toolkits and methodologies for explainability, bias mitigation, and robustness, which are essential for fostering trust in AI systems. Additionally, this paper examines challenges throughout the AI pipeline, from data collection to model deployment, and concludes with recommendations and research questions aimed at addressing these issues. By offering actionable insights, this study aims to guide researchers, practitioners, and policymakers in developing ethical and reliable AI systems that align with the principles of Industry 5.0, ensuring both technological advancement and societal value.



Academic Editor: Tao Peng

Received: 19 May 2025

Revised: 18 June 2025

Accepted: 8 July 2025

Published: 11 July 2025

Citation: Ahangar, M.N.; Farhat, Z.A.; Sivanathan, A. AI Trustworthiness in Manufacturing: Challenges, Toolkits, and the Path to Industry 5.0. *Sensors* **2025**, *25*, 4357. <https://doi.org/10.3390/s25144357>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: Artificial Intelligence (AI); manufacturing; Industry 4.0; Industry 5.0; AI trustworthiness; transparency; fairness; robustness; accountability; ethical AI; bias mitigation; explainability; AI Toolkits; sustainable manufacturing; human-centric AI

1. Introduction

The exponential growth of digital systems in recent years has led to the generation of large-scale, high-dimensional data. This trend is particularly evident in modern manufacturing, where a key shift has been the adoption of decentralised and distributed architectures. In such systems, control and decision-making responsibilities are shared across multiple autonomous units, rather than being managed centrally. This decentralised structure enhances flexibility and resilience, enabling factories to respond swiftly to disruptions and sustain operations despite localised failures. While this distributed approach is often presented as a technical improvement, it also raises critical questions about coordination, accountability, and the reliability of decision-making across autonomous units—issues that are central to the trustworthiness of digital manufacturing systems. Moreover, the distribution of control

supports scalable and adaptable production models suited to high variability in demand and complexity [1].

However, this structural evolution also introduces significant challenges. The proliferation of interconnected devices, sensors, and machines leads to massive volumes of heterogeneous data being generated in real time. Traditional analytical tools and human-led analysis are increasingly inadequate for extracting actionable insights from this data deluge. As a result, more advanced, automated, and context-aware data processing methods are required to support intelligent decision-making in distributed manufacturing environment. As a result, Artificial Intelligence (AI) has become essential for intelligent data acquisition, management, and processing [2]. AI enables organisations to analyse large datasets, extract meaningful insights, and support informed decision-making. Efficient data management, supported by AI, not only enhances scalability, security, and operational efficiency but also minimises resource consumption [3]. In this review, we use the term AI to refer specifically to computational systems that can perform tasks typically requiring human intelligence, with a particular emphasis on learning from data and making decisions in complex environments [4]. The concept of knowledge in AI can be categorised into four types: definitional (explicit definitions and facts), deductive (logical inference from rules), inductive (generalisation from examples), and creative (generation of novel ideas) [5]. In this study, we focus on AI as knowledge derived from complex induction, encompassing machine learning, deep learning, and related data-driven approaches. This scope does not include all possible forms of AI, such as purely rule-based or symbolic systems, and is limited to the opportunities and challenges of data-driven, inductive AI in manufacturing environments.

To address the analytical challenges posed by decentralised and data-intensive manufacturing systems, AI and related digital technologies are increasingly employed to create high-fidelity digital models—often referred to as digital twins—that simulate real-world operations. The adoption of digital twins and context-aware processing is not merely a matter of technological advancement; their effectiveness and acceptance depend fundamentally on the trust stakeholders place in the underlying AI systems. Without explicit mechanisms for transparency, fairness, and accountability, these advanced tools risk introducing new vulnerabilities or amplifying existing biases. These models enable manufacturers to evaluate decision scenarios in a virtual environment, assess the impact of process changes, and prioritise risk mitigation strategies. By proactively identifying and responding to operational challenges, factories can improve resilience, reduce economic disruptions, and seize emerging opportunities with greater agility.

Building on the foundation established by Industry 4.0, the emerging paradigm of Industry 5.0 places greater emphasis on human-centric approaches, sustainability, and ethical considerations. In this new era, AI continues to play a pivotal role—not only in driving automation and efficiency but also in supporting more responsible and inclusive industrial practices [6,7]. To further conceptualise the flow and use of information in Industry 4.0 and 5.0, it is helpful to consider three fundamental components: Syntax, Semantics, and Pragmatics. Syntax refers to the structure and format of data, ensuring interoperability between systems. Semantics addresses the meaning and interpretation of data, enabling both machines and humans to derive actionable insights. Pragmatics, however, concerns the practical application and real-world impact of information—how data-driven outputs are used in operational contexts. Critically, trustworthiness is a central aspect of Pragmatics, as it determines whether stakeholders can reliably act on AI-generated insights in manufacturing environments. This perspective highlights that trustworthiness is not merely a technical attribute but a practical necessity for the successful and responsible adoption of AI in Industry 4.0/5.0 [8]. However, the adoption of AI in manufacturing is not without significant challenges. One major concern is the “black box” nature of many

AI models, which refers to the difficulty in understanding how these systems arrive at their decisions or predictions. This lack of transparency and interpretability can hinder trust and accountability, as stakeholders may be unable to trace or justify the reasoning behind AI-driven outcomes [9,10].

Trust in AI, particularly in high-stake industrial contexts, is a multidimensional construct. Drawing from the broader trust literature, trust can be understood as comprising three interrelated components: scientific or technical competence, effective communication, and shared values [11]. While much of the AI literature focuses on technical robustness and explainability (science/competence), empirical research consistently finds that failures of trust are more often rooted in value misalignments and poor communication than in technical shortcomings. As Greenberg [11] notes, value-based trust is often the most challenging to build and maintain, especially when organisational or societal values are perceived to be at odds with those of affected stakeholders.

Another critical issue involves biases present in both the data used to train AI systems and the algorithms themselves. Biases can arise from historical data that reflect existing inequalities or from the design of algorithms that inadvertently favour certain groups or outcomes over others. In manufacturing, such biases may result in unfair resource allocation, exclusion of certain workforce segments, or suboptimal decision-making that does not account for the diversity of real-world scenarios. These concerns threaten the fairness and inclusivity of AI applications, making it essential to identify, measure, and mitigate bias throughout the AI lifecycle [9].

In this study, AI trustworthiness is defined as the degree to which AI systems can be relied upon to operate transparently, fairly, robustly, and accountably within manufacturing environments. Drawing on established frameworks such as the European Commission's Ethics Guidelines for Trustworthy AI, we operationalise AI trustworthiness through its core dimensions. Transparency refers to the extent to which AI decision-making processes are understandable and explainable to stakeholders, enabling traceability and auditability. Fairness is the assurance that AI systems do not propagate or amplify bias and that outcomes are equitable across different groups and contexts. Robustness signifies the resilience of AI systems to errors, adversarial attacks, and changing operational conditions, ensuring reliable performance. Accountability denotes the presence of mechanisms for assigning responsibility and enabling recourse in the event of system failures or unintended consequences. In the manufacturing domain, these dimensions are particularly salient due to the high stakes associated with safety, quality, and regulatory compliance. Measurable characteristics of AI trustworthiness in this context include the availability of model documentation, bias detection and mitigation reports, robustness testing results, and clear lines of responsibility for AI-driven decisions [12].

The motivation for this paper stems from the urgent need to address these challenges and bridge the gap between rapid technological advancements and their ethical, human-centric application in manufacturing. Furthermore, the dynamic and complex nature of manufacturing environments amplifies the risks associated with AI failures, which can lead to significant supply chain disruptions, reduced business efficiency, loss in production capacity, ethical violations, and loss of stakeholder trust. Despite a growing body of research on AI in manufacturing, there remains a critical gap in comprehensively addressing the trustworthiness of AI systems. This gap is particularly significant in the context of Industry 5.0, where aligning AI technologies with human-centric and sustainable principles is paramount. To address these challenges, a range of organisations and regulatory bodies have established comprehensive frameworks to promote the trustworthy and responsible use of AI. Notable examples include the European Commission's Ethics Guidelines for Trustworthy AI, the National Institute of Standards and Technology (NIST) AI Risk Manage-

ment Framework, the Organisation for Economic Co-operation and Development (OECD) Principles on Artificial Intelligence, the International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC) standards on AI trustworthiness, the IEEE Ethically Aligned Design, and the Singapore Model AI Governance Framework. These frameworks provide structured approaches for organisations to assess, monitor, and improve the reliability, fairness, and ethical alignment of AI systems [13,14]. However, these frameworks differ in their scope, rigour, and practical enforceability. For example, while the European Commission's guidelines emphasise ethical principles, the NIST and ISO/IEC standards focus more on technical and procedural aspects. Contradictions and gaps remain, particularly regarding how these frameworks address the unique operational realities of manufacturing, such as real-time decision-making and the integration of legacy systems [15].

Nevertheless, there remains a critical research gap: the absence of comprehensive frameworks and methodologies specifically tailored to the unique demands of manufacturing environments. This paper aims to address this gap by providing a comprehensive survey of AI trustworthiness in manufacturing. The primary objectives of this study are the following:

1. Critically examine the role of AI in the transition from Industry 4.0 to Industry 5.0, with a focus on the technical, ethical, and organisational challenges specific to manufacturing.
2. Assess the effectiveness and limitations of existing toolkits for ensuring AI trustworthiness—specifically transparency, fairness, robustness, and accountability in manufacturing contexts.
3. Formulate targeted research questions and methodological approaches to address the most pressing challenges of AI adoption in manufacturing, drawing on industry case studies

The findings of this study are expected to guide professionals, engineers, and decision-makers in manufacturing to adopt AI in ways that improve processes and respect societal and environmental values.

The structure of this paper, illustrated in Figure 1, is organised into eight sections, each addressing a critical dimension of AI trustworthiness in the context of Industry 5.0. Section 2 provides the background, tracing the evolution from Industry 4.0 to Industry 5.0 and emphasising the shift toward human-centricity, sustainability, and ethical integration. It further details the collaborative nature of Industry 5.0 and enumerates practical AI use cases in manufacturing, such as digital twins, predictive maintenance, and generative design. This section sets the stage for addressing the first research question: What is the role of AI in the transition from Industry 4.0 to Industry 5.0, and what are the associated challenges? Section 3 outlines the multi-stage methodology, encompassing a literature review, systematic search and selection, toolkit analysis, and an interdisciplinary, human-centric approach to synthesising research questions. This section establishes the methods used to explore the research questions and address AI adoption challenges in manufacturing. Section 4 examines the challenges in AI adoption for Industry 5.0, including technical, organisational, and ethical barriers like black-box models, data bias, reliability, regulatory concerns, security threats, and workforce adaptation. This section directly addresses the second part of the first research question: What are the technical, ethical, and organisational challenges of AI adoption in manufacturing? Section 5 discusses the importance of AI trustworthiness, drawing on real-world failures to highlight the necessity of transparency, fairness, and accountability in building trust and preventing harm. This section underscores the need for trustworthy AI systems and sets the context for exploring the necessary toolkits. Section 6 defines the core factors of AI trustworthiness: explainability, accountability, fairness, and robustness. It reviews key toolkits and frameworks for each factor, addressing the need for interpretable decisions, responsible governance, context-specific fairness, and resilience against errors and adversarial attacks. This section directly addresses the

second research question: What toolkits are necessary to ensure AI trustworthiness, such as transparency, fairness, robustness, and accountability? Section 7 analyses challenges across the AI pipeline, from data collection and preprocessing to model development and deployment, and formulates research questions on data integrity, bias detection, labelling consistency, and the trade-offs between interpretability and performance. This section delves into the methods to address AI adoption challenges in manufacturing, supported by industrial examples, thus addressing the third research question. Section 8 concludes by summarising the progress in trustworthy AI frameworks and toolkits, underscoring the ongoing need for ethical, technical, and organisational vigilance. It calls for interdisciplinary collaboration, regulatory compliance, and practical evaluation of toolkits in real-world scenarios and outlines future research directions to ensure continuous monitoring and adaptation of AI systems in manufacturing. This section synthesises the findings and provides a roadmap for future research, addressing all three research questions.

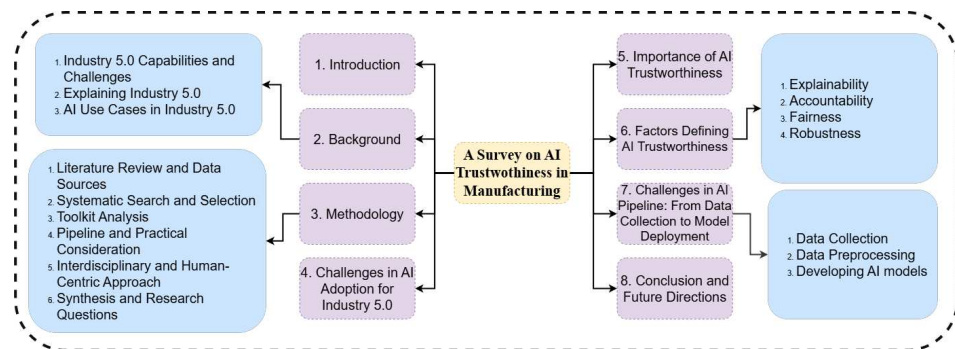


Figure 1. Overall paper structure.

2. Background

2.1. Industry 5.0 Capabilities and Challenges

The evolution of modern manufacturing began with Industry 4.0, which is characterised by the integration of cyber–physical systems, the Internet of Things (IoT), and advanced data analytics into industrial processes. Industry 4.0 has enabled unprecedented levels of automation, connectivity, and data-driven decision-making, transforming traditional factories into smart, interconnected environments. However, this transformation has also introduced significant challenges, such as managing the complexity of large-scale data, ensuring cybersecurity, and addressing the skills gap required to operate and maintain advanced technologies.

Figure 2 outlines the key pillars driving Industry 4.0, as identified in [16]. The influence of these technologies now extends well beyond traditional industrial settings, shaping home products, business models, clean energy solutions, and broader sustainability efforts—areas that earlier industrial revolutions largely overlooked. As a result, industry is increasingly recognised as a catalyst for systemic transformation, pushing economies toward greater sustainability [17]. Achieving this shift requires the integration of societal and environmental considerations as core priorities within the industrial sector.

This move toward decentralisation has been made possible by the widespread adoption of sensors and actuators embedded in machines through the IoT. These devices create seamless connectivity with computing systems and generate vast streams of data, commonly referred to as Big Data [18]. To handle this data efficiently, processing often occurs locally on IoT devices or is distributed through cloud and edge computing platforms. This approach not only optimises costs and improves scalability by leveraging virtual resources [19] but also supports the adoption of new technologies aligned with Industry 4.0 objectives. AI technologies are uniquely capable of rapidly collecting and analysing

information from multiple systems, enabling tasks such as fault prediction and action selection to be performed with far greater efficiency [20]. Consequently, many companies are adopting intelligent systems that support various levels of process automation, further accelerating the transformation of the manufacturing industry and supporting the broader goals of Industry 4.0 and beyond.

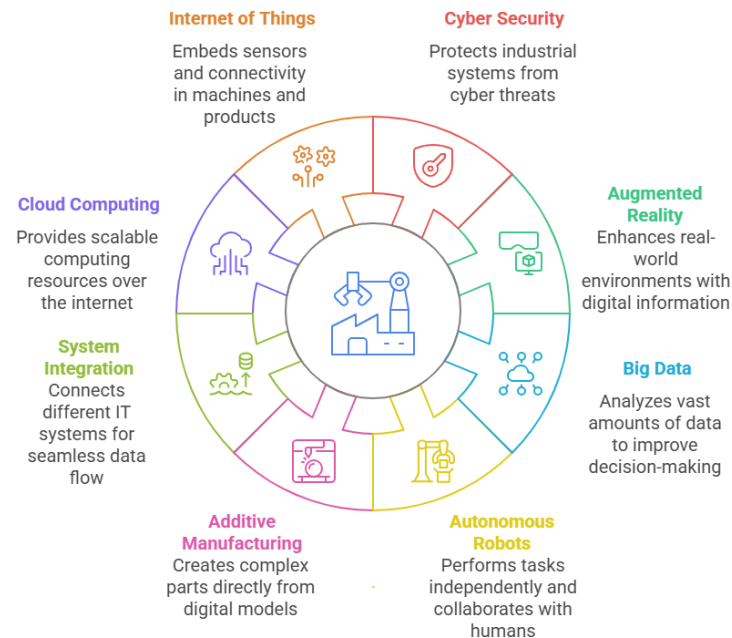


Figure 2. Key technologies in Industry 4.0.

The European Commission introduced the concept of Industry 5.0 in 2020 during a dedicated workshop involving research and technology organisations and funding bodies. This new paradigm integrates AI and the societal dimension as key drivers for the future of European industry [21]. Since then, multiple initiatives have been launched to support Industry 5.0, including efforts to upskill and reskill European workers, particularly in digital competencies (Skills Agenda and Digital Education Action Plan); fostering a more competitive industrial landscape through accelerated investment in research and innovation (Industrial Strategy); promoting sustainable development through resource-efficient, eco-friendly industries and a transition to a circular economy (Green Deal); and advocating for a human-centric approach to digital technologies via regulatory frameworks such as the AI Act, white papers, and trustworthy AI requirements [22–27].

A central pillar of Industry 5.0 is AI adoption, with a focus on high-speed data processing, workforce expertise in managing AI-driven heterogeneous technologies (including computing resources and data), and embedding ethical principles throughout the AI lifecycle to ensure trust and safe working environments [28,29]. While Industry 5.0 aims to foster collaboration between humans and machines and promote ethical, sustainable industrial practices, it also faces its own set of challenges. These include integrating ethical principles into AI systems, ensuring workforce adaptability, and balancing technological progress with societal and environmental considerations. While Industry 5.0 is often presented as a progressive and human-centric evolution of manufacturing, several of its core assumptions warrant critical examination. For example, the notion that increased human–machine collaboration will automatically lead to more ethical or sustainable outcomes is not universally supported by empirical evidence. There is ongoing debate about whether the integration of advanced AI and automation truly empowers workers or, conversely, risks further deskilling and job displacement. Additionally, the emphasis on sustainability and resilience

in Industry 5.0 frameworks can sometimes mask the persistent tension between economic growth and environmental limits, raising questions about the feasibility of achieving all three goals simultaneously. Critics also point out that the practical implementation of ethical AI principles remains challenging, with many organisations struggling to translate high-level values into operational practices. As such, while Industry 5.0 offers an aspirational vision, its real-world impact depends on addressing these unresolved tensions and ensuring that technological progress is matched by genuine social and environmental responsibility [7].

Despite ongoing research efforts to integrate ethical considerations into AI applications, unique challenges persist depending on the operational environment and the specific domains where these technologies are deployed [30].

2.2. Explaining Industry 5.0

In the context of Industry 5.0, the connection between factory-specific challenges, AI applications, and technological pillars remains insufficiently defined. Industry 5.0 marks a fundamental shift that extends beyond technological and economic aspects, placing a strong emphasis on human well-being, sustainability, and circular economies. Unlike previous industrial advancements that focused on automation and efficiency, Industry 5.0 promotes a collaborative relationship between humans and machines, leveraging their unique strengths rather than aiming for human replacement [31]. This paradigm shift calls for a more holistic integration of AI, where ethical considerations and societal impacts are prioritised alongside technical advancements.

While Industry 5.0 aspires to be human-centric and ethical, these ambitions often involve complex trade-offs and can lead to unintended consequences. For example, efforts to enhance worker well-being through increased human-machine collaboration may inadvertently introduce new forms of workplace stress, such as the need for constant upskilling or the psychological impact of working alongside intelligent machines. Similarly, prioritising ethical AI can sometimes slow down innovation or increase operational costs, as organisations must invest in transparency, bias mitigation, and compliance measures. There is also the risk that well-intentioned ethical frameworks may be inconsistently applied, leading to gaps between policy and practice. These examples highlight that the pursuit of human-centric and ethical objectives in Industry 5.0 is not without challenges, and careful consideration of potential trade-offs is essential for responsible implementation [32].

In addition to these user-facing applications, AI is increasingly being deployed within the underlying infrastructure, such as edge computing. Here, AI enables real-time data processing for tasks like predictive maintenance and supports advanced features, including Augmented Reality experiences. This seamless integration of AI across both consumer applications and technical infrastructure demonstrates its versatility and growing importance in modern technology ecosystems [33]. As AI continues to evolve, its role in shaping both the digital landscape and industrial environments will become even more significant, underscoring the need for ongoing research and thoughtful implementation.

Several key technologies serve as enablers of Industry 5.0, as identified by authors in Xu and Duan [2], Xu et al. [7], Commission et al. [21], Wang et al. [31], Habib ur Rehman et al. [34], Vyhmeister et al. [35,36], Wang et al. [37], Wu et al. [38]. These enablers, outlined in Figure 3, represent core components that drive this new industrial paradigm [39].



Figure 3. Key technologies in Industry 5.0.

Future industries are expected to play an important role in advancing societal goals while contributing to a more environmental friendly and sustainable ecosystem [40].

2.3. AI Use Cases in Industry 5.0

AI plays a significant role in Industry 5.0 by enabling smarter, more efficient, and adaptable operations. The following examples shown in Figure 4 illustrate AI's potential applications within the Industry 5.0 framework [16,41,42]:

1. **Digital Twin:** AI is utilised to create virtual representations of processes, production systems, factories, and supply chains, referred to as digital twins. These virtual models are employed to simulate, evaluate, and predict performance in real-time. By replicating the physical environment, digital twins allow manufacturers to monitor and improve operations without needing direct engagement with the physical assets. They depend on data from IoT sensors, programmable logic controllers (PLCs), deep learning techniques, and AI algorithms to continuously update the digital model with real-time information, ensuring an up-to-date and accurate virtual replica.
2. **Predictive maintenance:** AI processes sensor data from machinery to predict potential failures before they happen. By utilising a digital twin to examine patterns in equipment behaviour and performance, these systems can notify operators of potential issues in advance, enabling them to prevent breakdowns before they worsen. For instance, automotive manufacturers use predictive maintenance on assembly-line robots, greatly decreasing unplanned downtime and leading to significant cost savings. This method also allows manufacturers to schedule maintenance during off-peak hours, minimising disruptions to production timelines [43].
3. **Custom Manufacturing:** AI empowers manufacturers to provide mass customisation, enabling products to be tailored to individual customer preferences without disrupting production speed. By incorporating AI into the design process, companies can swiftly adjust designs in response to real-time consumer feedback. For example, clothing manufacturers utilise AI algorithms to personalise products, allowing customers to select designs that align with their unique tastes. This adaptability not only improves customer satisfaction but also boosts engagement by offering a more personalised shopping experience.
4. **Generative Design:** This technology allows manufacturers to explore numerous design possibilities by considering factors like materials and manufacturing limitations. This approach accelerates the design process by enabling the rapid evaluation of multiple iterations. Generative AI design tools are already being utilised in industries like the aerospace and automotive industries, where companies use them to develop

- optimised parts. Although the technology is already in use, its complete potential is still not being explored within the dynamic landscape of modern manufacturing.
5. **Quality Control:** AI improves quality control by using computer vision and machine learning, often supported by a digital twin, to detect defects in real-time. These systems examine product images during the manufacturing process, identifying inconsistencies or faults with greater precision than human inspectors. For example, electronic manufacturers utilise AI-driven quality control to ensure components meet stringent specifications. This results in higher product quality, reduced waste, and greater customer satisfaction.
 6. **Supply Chain Management:** AI streamlines supply chain operations by analysing large volumes of data to forecast demand, manage stock levels, and improve logistics. When coupled with a digital twin, AI can build a virtual model of the entire supply chain, enabling manufacturers to predict and simulate disruptions or shortages in real-time. Machine learning assists with demand predictions and automates procurement, ensuring that manufacturers receive materials precisely when needed. AI-driven order management systems also optimise order fulfilment, ensuring deliveries are made on time. For example, food manufacturers use AI to anticipate seasonal shifts in demand, allowing them to better manage resources and reduce waste. This ultimately boosts operational efficiency and enhances responsiveness to market fluctuations.
 7. **Inventory management:** AI enhances inventory management by analysing data to predict stock requirements and streamline replenishment. By forecasting demand and tracking inventory in real-time, manufacturers can ensure optimal stock levels, lowering storage costs and improving cash flow. For instance, food and beverage manufacturers use AI systems to monitor ingredient consumption as it happens. This enables them to predict future needs based on production timelines, seasonal factors, and historical usage, helping to avoid production disruptions and minimise waste from excess stock.
 8. **Energy Management:** AI systems track energy consumption in real-time to pinpoint inefficiencies. These systems can suggest changes that help cut energy costs and reduce environmental impact. For example, electronic manufacturers use AI-driven energy management solutions to improve their operations, leading to substantial cost reductions and a smaller carbon footprint.

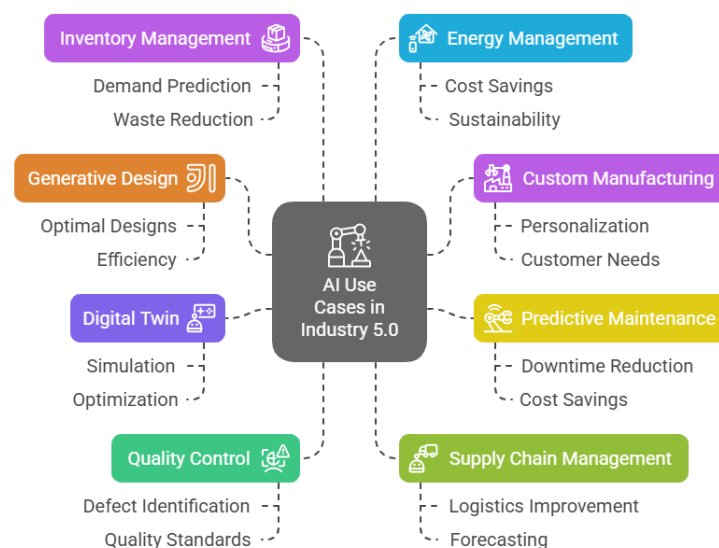


Figure 4. AI use cases in Industry 5.0.

While these AI applications offer significant promise, their deployment in manufacturing has also revealed critical challenges related to trust, fairness, and explainability. In manufacturing, “ethical” AI refers to systems that operate transparently, avoid bias, respect stakeholder values, and ensure accountability for outcomes [12]. However, real-world failures highlight the complexity of achieving these goals. Given the critical role of AI in Industry 5.0, industries are increasingly cautious about its adoption due to concerns over transparency, ethical risks, regulatory compliance, and reliability. Without clear governance and accountability, AI adoption remains a challenge, particularly in high-risk sectors.

3. Methodology

This study adopts a rigorous, multi-stage methodology to systematically investigate the challenges and enablers of AI trustworthiness in manufacturing, particularly within the context of evolving industrial paradigms. The approach highlighted in Figure 5 is designed to ensure both breadth and depth, combining a comprehensive literature review, critical analysis of toolkits, and the use of illustrative case studies to provide a holistic understanding of AI trustworthiness.

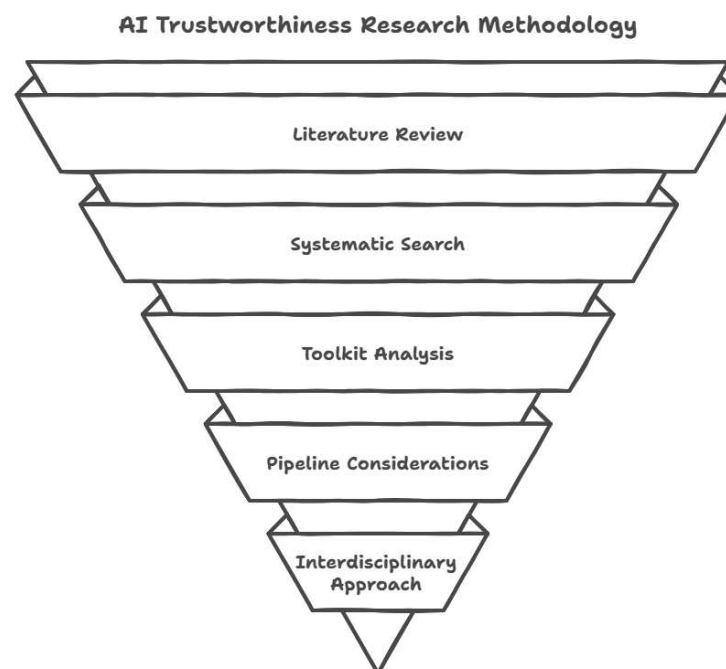


Figure 5. Overall methodology.

1. **Literature Review and Data Sources:** This research begins with an extensive literature review, targeting peer-reviewed journal articles, conference proceedings, and authoritative industry reports. Sources are drawn from high-impact databases including IEEE Xplore, Scopus, Web of Science, and the ACM Digital Library. To ensure relevance and currency, this review is limited to works published within the last decade, with a particular emphasis on studies addressing AI trustworthiness in manufacturing. In addition, regulatory documents and guidelines—such as ISO/IEC standards, the European Union (EU) AI Act, and the Assessment List for Trustworthy Artificial Intelligence (ALTAI)—are included to capture the evolving landscape of ethical and legal requirements.
2. **Systematic Search and Selection:** A systematic search strategy is employed, using targeted keywords such as “AI trustworthiness”, “Industry 5.0”, “ethical AI”, “transparency in AI”, “Toolkits in AI”, and “manufacturing.” The selection process prioritises studies that address the core dimensions of trustworthy AI—transparency,

fairness, robustness, and accountability—within manufacturing contexts. The inclusion of case studies, both of AI successes and failures, provides practical grounding and validation for the findings. To ensure a comprehensive and transparent literature review, the search strategy involved querying databases using specific search strings like “AI trustworthiness” AND “manufacturing”, “AI ethics” AND “smart manufacturing”, and “responsible AI” AND “industry 5.0”. Inclusion criteria were applied to select peer-reviewed articles, conference papers, and relevant reports published between 2015 and 2024, focusing on AI trustworthiness in manufacturing applications. Exclusion criteria were used to filter out studies that were not directly relevant to the manufacturing sector or did not address AI trustworthiness. The initial search yielded 500 of articles, which were then screened based on their titles and abstracts. Full-text reviews were conducted on 300 articles, resulting in a final selection of 200 articles that met the inclusion criteria.

3. **Toolkit Analysis:** The analysis is structured around four key dimensions of AI trustworthiness: transparency, fairness, robustness, and accountability. For each dimension, this study critically examines a range of prominent toolkits and frameworks, including but not limited to AI Explainability 360 (AIX360), SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), AI Fairness 360 (AIF360), FairLearn, IBM Adversarial Robustness Toolbox (IBM ART), and CleverHans. The discussion evaluates the strengths, limitations, and practical applications of these tools, offering a comprehensive perspective on their contributions to trustworthy AI in manufacturing. The evaluation criteria included: (1) transparency mechanisms (e.g., explainable AI (XAI) techniques), (2) fairness metrics and mitigation strategies, (3) robustness testing and validation methods, (4) accountability frameworks, and (5) ethical guidelines and compliance support. The toolkits were assessed based on their functionalities, ease of use, and applicability to manufacturing contexts. The evaluation involved a qualitative, comparative analysis, drawing upon expert judgment to assess the toolkits’ strengths and weaknesses in addressing AI trustworthiness concerns. A formal numerical scoring system was not used due to the diversity of toolkit functionalities and the context-dependent nature of manufacturing applications. Instead, the evaluation focused on providing a nuanced understanding of each toolkit’s capabilities and limitations in promoting AI trustworthiness.
4. **Pipeline and Practical Considerations:** The methodology explicitly addresses challenges across the entire AI pipeline—from data collection and preprocessing to model training, deployment, and post-deployment monitoring. Special attention is given to issues such as data quality, interoperability, bias, concept drift, and the integration of domain expertise.
5. **Interdisciplinary and Human-Centric Approach:** Recognising the complexity of manufacturing environments, the methodology emphasises interdisciplinary collaboration among AI developers, domain experts, and end-users. This ensures that technical solutions are both practically relevant and ethically aligned. The approach is further informed by the human-centric and sustainable ethos of Industry 5.0, integrating ethical considerations and stakeholder perspectives at every stage. No new stakeholder interviews or primary qualitative data were collected; instead, the human-centric perspective is embedded through the integration of ethical considerations and stakeholder insights from published qualitative studies and documented experiences. To illustrate the potential challenges and implications of AI trustworthiness in manufacturing, this study employs a series of hypothetical case studies. These scenarios are not based on specific real-world implementations but are carefully constructed to represent common AI applications across diverse manufacturing sectors such as

automotive, aerospace, and electronics. The purpose of these illustrative cases is to explore potential issues related to transparency, fairness, robustness, and accountability that could arise when deploying AI solutions in these contexts. By analysing these hypothetical scenarios, this study aims to provide insights into the proactive measures and strategies that manufacturing organisations can adopt to ensure AI trustworthiness.

6. **Limitations and Bias Mitigation Strategies:** As with any research, this study is subject to certain limitations. To address potential biases, several mitigation strategies were implemented throughout the research process. The possibility of selection bias in the illustrative case studies was reduced by ensuring a diverse representation of manufacturing sectors and AI application areas. To mitigate publication bias in the literature review, both peer-reviewed articles and grey literature sources (e.g., industry reports; white papers) were considered. The analytical subjectivity inherent in the toolkit evaluation was addressed through the use of a structured evaluation framework, clear evaluation criteria, and the involvement of multiple researchers in the analysis process to promote inter-rater reliability. While these strategies do not eliminate bias entirely, they significantly reduce its impact on this study's findings.
7. **Synthesis and Research Questions:** Findings from the literature, toolkit evaluations, and case studies are synthesised to identify persistent gaps and emerging best practices. This study formulates open research questions to guide future inquiry, particularly regarding the operationalisation of trustworthy AI in dynamic, real-world manufacturing settings.

By combining a systematic review, critical analysis, and practical validation, this methodology aims to advance the understanding and implementation of trustworthy, ethical, and human-centric AI systems in manufacturing, supporting the broader objectives for manufacturing. The subsequent section examines the major technical, organisational, and ethical challenges that hinder AI adoption in Industry 5.0 manufacturing. It discusses issues such as black-box models, data quality, reliability, regulatory uncertainty, and workforce adaptation.

4. Challenges in AI Adoption for Industry 5.0

AI empowers the manufacturing industry to adapt to changing market demands, personalise products at scale, and strengthen supply chain resilience through advanced data analytics and automation. However, its successful integration into Industry 5.0 is not without challenges. As shown in Figure 6, various technical, organisational, and ethical barriers must be addressed to ensure AI's seamless adoption and long-term impact [44].

1. **Technical Challenges:** A primary technical challenge is the "black box" nature of many AI models, which lack transparency and make it difficult for operators to trust or verify their decisions, raising concerns about accountability. The European Commission's Ethics Guidelines stress the need for explainable and transparent AI to foster user trust [45,46]. Another issue is the shortage of high-quality, relevant data for training AI models. Poor or biased data can lead to inaccurate results and reinforce existing biases, limiting AI's effectiveness in manufacturing [47]. Reliability is also a concern, as AI models that perform well in controlled settings may not replicate their success under real-world conditions due to variations in data distributions and unforeseen operational challenges, leading to inconsistent performance and affecting production quality and efficiency [48].
2. **Security and Cybersecurity:** Security concerns are paramount, as AI systems are susceptible to cyber threats that can compromise sensitive industrial data and disrupt operations. For example, adversarial attacks on machine learning models can

- manipulate outputs or cause system failures. The NIST Cybersecurity Framework and ISO/IEC 27001 provide standards for securing industrial AI systems, but their implementation in dynamic manufacturing environments remains challenging [49,50].
3. **Ethical and Regulatory Challenges:** Ethical considerations further complicate AI adoption. The potential for AI systems to perpetuate biases or make decisions that lack fairness necessitates the development of robust ethical frameworks. The European Commission’s guidelines advocate for AI that is lawful, ethical, and robust, ensuring adherence to principles such as fairness, accountability, and respect for privacy [45]. Regulatory uncertainty is a significant barrier, particularly where existing regulations conflict with AI optimisation. For instance, the General Data Protection Regulation (GDPR) mandates the right to explanation for automated decisions, which can conflict with the opacity of some machine learning models. This tension between data privacy and model transparency creates compliance challenges for manufacturers seeking to deploy advanced AI solutions [51]. The absence of standardised regulations and governing bodies for AI in manufacturing further exacerbates uncertainty, making it difficult for companies to ensure compliance and align with best practices. The European Commission’s ALTAI aims to provide actionable guidance to address these issues [52].
 4. **Organisational and Workforce Barriers:** The human-centric approach of Industry 5.0 emphasises the importance of collaboration between AI systems and human workers. Bridging the skills gap through targeted education and training programs is vital to equip the workforce with the necessary competencies to effectively interact with AI technologies [53]. Organisational resistance to change, lack of digital maturity, and insufficient leadership support can also hinder successful AI adoption.
 5. **Barriers for SMEs versus Large Manufacturers:** Small and medium-sized enterprises (SMEs) face unique barriers compared to large manufacturers. SMEs often lack the financial resources, technical expertise, and access to high-quality data required for effective AI implementation. The cost of acquiring, integrating, and maintaining AI systems can be prohibitive, and SMEs may struggle to attract or retain skilled personnel. In contrast, large manufacturers typically have greater capacity to invest in digital infrastructure, data management, and workforce development, enabling them to overcome many of these barriers more readily. As a result, the digital divide between SMEs and large enterprises may widen, limiting the broader impact of AI in the manufacturing sector [54].



Figure 6. AI adoption challenges in Industry 5.0.

5. Importance of AI Trustworthiness

AI systems are revolutionising various aspects of life, from recommending movies to diagnosing illnesses, assisting customers, and much more [55]. While AI offers a vast range of applications, its rapid advancement has also sparked significant concerns. The late Stephen Hawking once warned that “If not properly regulated, AI has the potential to become the greatest threat to humanity” [56].

Today, AI plays a crucial role in decision-making across multiple industries, but its outcomes are not always favourable. The growing reliance on AI brings a significant responsibility to ensure that these systems do not cause harm to humanity. However, there have been instances where AI has failed, leading to severe consequences. For example, the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm, widely used in the United States (US) to predict criminal recidivism risk, was found to exhibit racial bias against Black individuals [57]. A facial recognition system misclassified Black people due to poor-quality training data [58]. Similarly, a major tech company’s AI-driven resume screening system displayed bias against women [59]. These cases illustrate how bias can distort the decisions of black-box AI models, leading to unfair and harmful outcomes.

In some situations, AI has even resulted in physical harm due to system failures. One such case involved a self-driving car that struck and killed a pedestrian because its algorithm malfunctioned and failed to respond correctly when detecting a person on the road [60]. Moreover, the complexity of AI models makes it difficult to interpret their decision-making process, limiting their adoption and effectiveness. For instance, the research [61] found that despite their potential benefits, AI-powered medical diagnosis support systems have seen limited adoption among healthcare professionals. This reluctance stems from the lack of interpretability in these systems, reducing doctors’ trust and willingness to use them. AI systems have now reached a level of performance that allows them to be widely integrated into society. These technologies are already reshaping people’s daily lives [62]. However, despite their usefulness, this does not automatically mean they are reliable or trustworthy. A casual approach toward AI is unacceptable, especially in high-risk applications where a single wrong decision can have severe consequences. These systems can be fragile and prone to bias.

Marcus and Davis [63] provide a compelling example using facial recognition technology to illustrate the necessity of trustworthy AI. If such software is used for automatically tagging individuals in social media photos, a lower degree of accuracy may be tolerable. However, the same system becomes unacceptable when employed by law enforcement to identify suspects from surveillance images. This contrast highlights how AI is more readily adopted when errors do not pose serious risks to individuals or society.

To maximise the benefits of AI in critical applications and encourage broader adoption, it is essential to understand the reason behind the decision taken by an AI system. The following section outlines key requirements necessary to ensure AI systems are safe, reliable, and trustworthy.

6. Factors Defining AI Trustworthiness

In recent years, numerous research institutions, private companies, and government bodies have introduced various frameworks and guidelines aimed at ensuring that AI is trustworthy [61,64–68]. However, the overwhelming number of proposed principles has made it challenging to establish a unified set of standards. To address this issue, some researchers [15,69] have analysed and compared these principles to identify areas of consensus. Their findings indicate an emerging agreement on five key principles: transparency/explainability, justice and fairness, non-maleficence (which includes societal

and environmental well-being), responsibility/accountability, and privacy. These principles appear more frequently in different frameworks compared to others.

To align with this analysis and adhere to one of the earliest government-backed AI frameworks, we have chosen the EU's framework for trustworthy AI [61], which incorporates all five principles while also emphasising the human-centred aspect of AI.

The EU outlined three core guidelines that AI systems should follow to be considered trustworthy: they must be lawful, ethical, and robust. Lawfulness ensures that AI development, deployment, and usage comply with existing regulations. Ethical considerations require AI to respect human values and moral principles. Robustness emphasises that AI must be technically reliable while also adhering to legal and ethical standards. These guidelines provide a foundational structure for developing and deploying AI responsibly.

To operationalise these guidelines and enhance AI trustworthiness, the EU [61], introduced four ethical principles, each supported by seven key requirements, as summarised in Figure 7 [70,71]. The first principle, respect for human autonomy, ensures AI complements human decision-making rather than replacing it. The second principle, prevention of harm, guarantees that AI functions as intended without causing unintended damage to individuals or society. The third principle, fairness, ensures AI systems treat all individuals and social groups equitably, without bias or discrimination. Lastly, the fourth principle, explainability, ensures AI systems remain transparent and interpretable. These principles are explained through the following key requirements, which align with the aforementioned ethical principles:

1. **Human Agency and Oversight:** AI systems should support and enhance human decision-making rather than replace it. Human involvement should be proportional to the risks and societal impact of AI's decisions [72,73].
2. **Technical Robustness and Safety:** AI systems must be reliable and function as intended. They should be capable of recovering from failures without harm and handle errors throughout the AI lifecycle. The system must also resist external threats and produce reproducible results [29].
3. **Privacy and Data Governance:** AI systems must safeguard user data throughout its lifecycle, ensuring compliance with data protection regulations like the General Data Protection Regulation (GDPR). Sensitive data must be protected from misuse [74].
4. **Transparency:** AI systems should be understandable, with decisions that can be explained, interpreted, and reproduced. Stakeholders should fully grasp the system's performance and limitations [75,76].
5. **Diversity, Non-discrimination, and Fairness:** AI systems must ensure fairness, treating all societal groups equally and avoiding any form of discrimination, whether direct or indirect [77].
6. **Societal and Environmental Well-being:** AI systems should not harm society or the environment during their development, operation, or use [61].
7. **Accountability:** AI systems must be capable of justifying their decisions. There should be mechanisms for assigning responsibility for both correct and incorrect outcomes, along with regular audits to prevent harm [78].

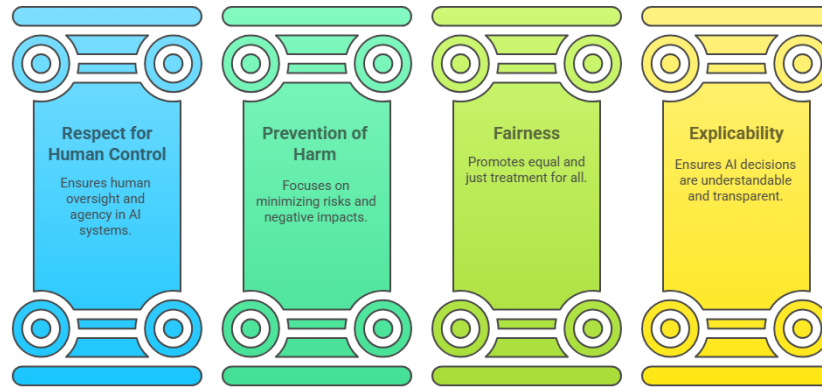


Figure 7. Trustworthy AI framework.

6.1. Explainability

Explainability is key to ensuring that the rationale behind AI-driven decisions is clear, supporting transparency and making the system easier to interpret. This can lead to system improvements and stronger governance practices [79–83].

AI systems that offer clear explanations help identify flaws and vulnerabilities, contributing to the overall trustworthiness of the system [84,85]. Users have the right to understand how an AI system produces results, including insight into the system’s decision-making process, the data used to train it, and the criteria used to evaluate its outcomes [86–88]. Additionally, AI systems should offer explanations that cater to a wide range of users, each with varying levels of expertise and specific needs [89]. When users comprehend the reasons behind an AI system’s decisions, their trust in the system increases [90]. It is important to recognise that explanations vary depending on their intended purpose and the user’s background, resulting in different approaches to interpretability, such as global and local interpretability [91,92]. Global interpretability aims to explain the overall workings of an AI system, providing a high-level view of how decisions are made. This type of interpretability is typically used in large-scale applications like climate modeling [93], where practical challenges arise due to its scale. On the other hand, local interpretability is more focused on explaining individual decisions made by the AI system, offering more immediate and context-specific insights. The timing and relevance of these explanations are determined by the data used and the stage of decision-making [94,95]. Local interpretability can be further divided into two types: ex ante and ex post explanations [96]. Ex ante explanations describe how the system works and is designed before it is used, ensuring that it is adequately tested and reliable. Ex post explanations, however, clarify the reasons behind decisions after they are made, validating the assumptions established by the ex ante explanations [97]. As outlined by ISO [98], both ex ante and ex post explanations are critical components of an AI system’s trustworthiness through transparency and interpretability.

Current methods for ensuring explainability primarily address the needs of developers and designers, aiding in debugging and oversight [99]. However, more suitable approaches are needed to address the needs of non-expert users, bridging the gap between transparency and actual implementation [100]. In this context, the toolkits highlighted in Figure 8 and explained in Table 1 provide a variety of approaches to AI explainability, each designed to tackle different aspects of model transparency. For example, AI Explainability 360 (AIX360) and Local Interpretable Model-agnostic Explanations (LIMEs) focus on providing local explanations and enhancing trust through user-friendly models, which can be particularly helpful for non-expert users who may not have a deep understanding of machine learning. On the other hand, Shapley Additive Explanations (SHAPs) offers both global and local

explanations that help users understand feature importance in a more comprehensive manner, which could be useful in settings requiring a higher degree of interpretability.

Table 1. Comparison of AI explainability toolkits.

Toolkit	Pros	Cons	Use Cases
AI Explainability 360 (AIX360) [101]	<ul style="list-style-type: none"> Comprehensive set of algorithms covering various explanation dimensions. Supports multiple data types, enhancing versatility. Developed by IBM, ensuring reliability and community support. 	<ul style="list-style-type: none"> Steep learning curve due to broad feature set. Some algorithms may require substantial computational resources. 	<ul style="list-style-type: none"> Understanding and interpreting predictions from complex machine learning models. Ensuring transparency and trustworthiness in AI-driven decision-making processes.
LIME (Local Interpretable Model-agnostic Explanations) [102]	<ul style="list-style-type: none"> Provides local explanations by approximating complex models with simpler ones. Model-agnostic; applicable to various machine learning models. Enhances user trust through understandable explanations. 	<ul style="list-style-type: none"> Local explanations may not fully capture global model behaviour. Performance can be affected by noisy data, leading to inconsistent interpretations. 	<ul style="list-style-type: none"> Explaining individual predictions in domains like healthcare and finance. Assisting in debugging and improving model performance by understanding specific decision paths.
SHAP (Shapley Additive Explanations) [103]	<ul style="list-style-type: none"> Offers both global and local explanations, providing a comprehensive view of model behaviour. Based on solid game-theoretic foundations, ensuring consistent and fair feature importance values. 	<ul style="list-style-type: none"> Computationally intensive, especially with large datasets and complex models. Requires careful handling of feature interactions to avoid misleading interpretations. 	<ul style="list-style-type: none"> Assessing feature importance in predictive models. Enhancing model transparency in sectors like finance and healthcare by elucidating the impact of individual features on predictions.
XAITK (Explainable AI Toolkit) [104]	<ul style="list-style-type: none"> Provides a suite of tools for analysing and understanding complex machine learning models. Includes analytics tools and methods for interpreting models, supporting various explanation techniques. 	<ul style="list-style-type: none"> May require integration efforts with existing workflows. Documentation and community support might be less extensive compared to more established toolkits. 	<ul style="list-style-type: none"> Analysing and interpreting complex machine learning models across various domains. Supporting research and development in AI transparency and accountability.
Quantus [105]	<ul style="list-style-type: none"> Offers a collection of evaluation metrics for assessing the quality of explanations. Facilitates the comparison of different explanation methods. Aids in identifying the most effective explanation techniques for specific models. 	<ul style="list-style-type: none"> Primarily focused on evaluating explanations rather than generating them. May require additional tools or methods for generating explanations. 	<ul style="list-style-type: none"> Evaluating the effectiveness of different explainability methods. Assisting researchers in selecting appropriate explanation techniques for their models.

Furthermore, concerns about privacy and security can deter organisations from adopting AI solutions. Therefore, approaches that guarantee explainability while safeguarding privacy and security must be carefully developed [106–109]. Some of these toolkits, such as Quantus and Explainable AI Toolkit (XAITK), focus on evaluating and ensuring the robustness of explanation methods, which could be crucial for addressing privacy concerns by ensuring the fairness and transparency of AI systems without exposing sensitive data.

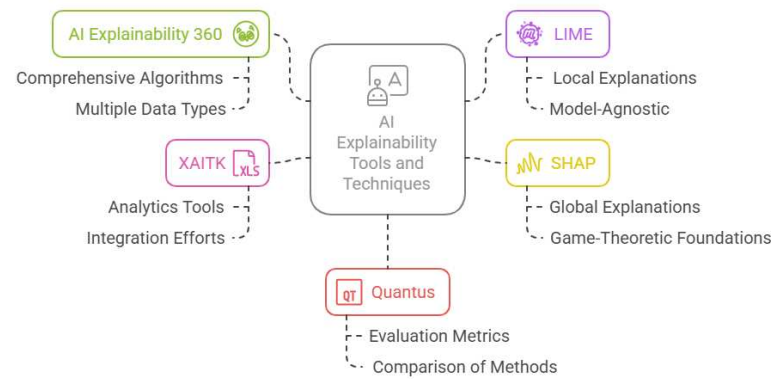


Figure 8. Explainability AI toolkits.

6.2. Accountability

To prevent algorithmic decision-making from leading to harmful outcomes, it is crucial to carefully oversee the design, deployment, and operation of these algorithms. Since algorithms are computer programs trained on data, those involved in their creation and user must take responsibility for any unintended consequences that arise [110–112]. In [78], the author characterises accountability as a collaborative effort, where different stakeholders are assigned responsibilities at various stages of the AI lifecycle. Essentially, ensuring accountability in algorithmic decision-making requires evaluating these systems against relevant standards and clearly defining the roles of those responsible for their development.

The increasing dependence on algorithmic decision-making (DM), particularly in high-risk environments, emphasises the need for strong accountability mechanisms. These algorithms must be designed, developed, and implemented in a reliable and secure manner to prevent potential failures. System malfunctions can have severe consequences, as demonstrated by the Boeing aircraft crash, which resulted in 346 fatalities due to software defects [113]. Similarly, Volkswagen encountered significant challenges with the software architecture of its electric vehicles, and a facial recognition system exhibited bias, disproportionately impacting women and individuals with darker skin tones [114]. Effective monitoring of these algorithms could help prevent such issues. However, assigning responsibility for these failures is complex—should the blame fall on developers, data collectors, or users trained to operate the system? The ambiguity surrounding accountability highlights the necessity for a well-structured framework [115,116].

Several strategies can enhance accountability in algorithmic DM. These include incorporating accountability measures into the algorithm's design, increasing transparency, and enforcing stringent regulations and policies to improve oversight. Since accountability is a dynamic process [117,118], establishing it requires comprehensive governance throughout the AI lifecycle and active collaboration among all stakeholders [119]. However, pinpointing liability in the event of a system failure is challenging, as multiple parties are typically involved in the development process. Accountability measures should be tailored to specific applications, as a universal framework may not be suitable for all domains. To enhance governance, it is recommended to implement context-specific accountability strategies [98]. For example, ISO standards address accountability in medical AI systems and AI-driven hiring tools. In medical AI, healthcare providers bear responsibility for any harm caused, as they are experts in their field and the system is intended only to support their decision-making. Conversely, in AI-based recruitment systems, users are not held accountable for negative outcomes since they lack insight into why their application was rejected. This distinction underscores the importance of designing accountability measures that align with the specific context of each application [120].

6.3. Fairness

AI-driven systems and algorithms process vast amounts of data and logical rules to perform specific tasks and support decision-making. Given the significant role these systems play in everyday activities and operations, it is crucial to ensure they function without bias. A fair AI system should not discriminate against any individual or societal group [77]. The concept of fairness is closely aligned with ethical principles and moral values [121–124].

When AI systems are designed, developed, implemented, or monitored unfairly, they can produce harmful outcomes. Numerous cases illustrate the consequences of biased AI. For example, a judicial system was found to incorporate a flawed risk assessment tool that disproportionately discriminated against individuals with darker skin tones [125]. Similarly, a prominent technology company faced scrutiny for using a biased hiring algorithm that disadvantaged women [126]. Research has also revealed that certain predictive analytic tools used in child maltreatment screenings unfairly discriminated against marginalised groups based on race and socioeconomic status [127]. Several factors influence the trustworthiness of AI, including biases in data, models, and evaluation processes. Given the critical importance of fairness in AI, various studies have attempted to define the concept, yet there is no universally accepted definition. Some researchers have analysed and compared different interpretations of fairness in AI [122]. Generally, fairness in AI is context-dependent, meaning its definition varies based on how and where AI is applied. The two primary categories of fairness in AI are individual fairness and group fairness [122]. Individual fairness ensures that individuals within the same category receive consistent predictions [128]. This concept is associated with fairness through awareness [128] or unawareness [129], as well as counterfactual fairness [130,131]. On the other hand, group fairness focuses on equitable treatment across different societal groups [132]. Various methods are used to evaluate fairness in AI, such as demographic parity, which ensures balanced representation across groups, equalised odds, which accounts for fairness in prediction outcomes, equal opportunity, which focuses on equitable access to favourable results, and conditional statistical parity, which adjusts fairness based on specific conditions [131,133–135].

Beyond defining fairness, various approaches have been developed to promote fairness in AI. However, identifying a single universal method to detect and eliminate all types of bias remains challenging. In [136] the authors highlighted the need for further research to explore different perspectives on fairness, particularly within AI applications, as certain systems may be more vulnerable to specific biases than others. Establishing comprehensive frameworks and policies that define fairness in AI based on application context is essential. Additionally, stakeholders may have differing interpretations of fairness, emphasising the need for inclusive discussions to enhance AI trustworthiness. Strengthening testing protocols and implementing effective measures to detect and mitigate bias in AI systems is also vital [137].

To effectively address bias and ensure fairness in AI systems, various toolkits have been developed to assess and mitigate discriminatory outcomes. These toolkits provide diverse methodologies for evaluating fairness, offering both technical and ethical approaches. The Tables 2 and 3 presents a comparative analysis of prominent AI fairness toolkits, highlighting their strengths, limitations, and application areas. Understanding these toolkits can be helpful for selecting the most suitable framework based on the specific needs of an AI system. Additionally, Figure 9 provides a visual representation of this comparison, further aiding in the evaluation and selection process.

Table 2. Summary of AI fairness toolkits (Part 1). These toolkits are used to assess and mitigate bias in AI systems, ensuring fairness in decision-making processes.

Toolkit Name	Advantages	Disadvantages	Use Cases
IBM AI Fairness 360 (AIF360) [138]	<p>Open-source toolkit for bias detection and mitigation.</p> <p>Provides a comprehensive set of metrics and algorithms for bias mitigation.</p> <p>Highly customisable and suitable for large-scale applications.</p>	<p>Focuses primarily on fairness, may not address other ethical AI dimensions.</p> <p>Can be complex to implement in production systems.</p> <p>May require significant computational resources.</p>	<p>Evaluating and mitigating bias in hiring, loan approvals, and other decision-making systems.</p> <p>Improving fairness in public services, education, and financial systems.</p> <p>Ensuring fairness in automated systems such as hiring and loan approvals.</p>
Microsoft FairLearn [139]	<p>Provides fairness assessment and bias mitigation tools.</p> <p>Supports multiple fairness metrics and mitigation strategies.</p> <p>Enables easy integration with scikit-learn models.</p>	<p>Requires technical expertise for implementation.</p> <p>Limited documentation for non-technical users.</p> <p>Limited flexibility in non-enterprise applications.</p>	<p>Fairness assessment in machine learning models, especially in enterprise settings.</p> <p>Ensuring fairness in predictive models used in hiring, finance, and healthcare.</p> <p>Used for bias mitigation in sensitive decision-making systems.</p>
Google What-If Tool [140]	<p>Interactive tool for exploring model predictions and fairness.</p> <p>Allows visual exploration of bias and fairness across model predictions.</p> <p>Easy-to-use interface for non-technical stakeholders.</p>	<p>Limited to TensorFlow models.</p> <p>Not suitable for non-TensorFlow based models.</p> <p>Can be time-consuming for large datasets.</p>	<p>Analyzing fairness in predictive models, such as fraud detection and medical diagnoses.</p> <p>Used for exploring fairness in machine learning models for fraud detection and healthcare applications.</p> <p>Assessing fairness in machine learning applications for public policy and social justice.</p>
Aequitas [141]	<p>Focuses on bias and fairness in decision-making systems. Designed to be used with real-world decision-making data.</p> <p>Strong documentation and user support.</p>	<p>Limited scope for other ethical AI principles.</p> <p>Does not provide technical tools for bias mitigation.</p> <p>Not highly customisable for complex AI systems.</p>	<p>Bias detection in public policy and social justice applications.</p> <p>Used for fairness assessments in hiring, criminal justice, and education systems.</p> <p>Ensuring fairness in decision-making systems related to government and social issues.</p>
ML Fairness Gym [142]	<p>Simulates long-term impacts of fairness interventions.</p> <p>Provides an interactive environment for testing fairness interventions.</p> <p>Supports a range of fairness interventions for experimentation.</p>	<p>Requires expertise in simulation modeling.</p> <p>High computational cost for large simulations.</p> <p>Simulation results may not generalise to all real-world scenarios.</p>	<p>Evaluating fairness in dynamic systems like credit scoring and hiring.</p> <p>Used for understanding fairness in long-term decision-making systems.</p> <p>Assessing fairness in evolving applications such as credit scoring and insurance.</p>

Table 3. Summary of AI fairness toolkits (Part 2). These toolkits are used to assess and mitigate bias in AI systems, ensuring fairness in decision-making processes.

Toolkit Name	Advantages	Disadvantages	Use Cases
AINow Algorithmic Impact Assessment Toolkit. [143]	<p>Engages stakeholders in assessing fairness and ethical implications.</p> <p>Provides a structured framework for ethical AI assessments.</p> <p>Emphasises the importance of human oversight in AI decision-making.</p>	<p>Limited technical tools for bias mitigation.</p> <p>Limited to ethical assessments rather than technical solutions.</p> <p>Lacks deep technical fairness metrics and algorithms.</p>	<p>Assessing fairness in community-focused AI applications.</p> <p>Used for impact assessments in AI systems affecting marginalised communities.</p> <p>Assisting in responsible AI implementation in public services.</p>
DotEveryone Consequence Scanning Toolkit	<p>Open-source; minimal resources required; focuses on societal impacts.</p> <p>Helps in identifying the societal consequences of AI deployments.</p> <p>Allows for early detection of ethical and social risks in AI systems.</p>	<p>Requires a strong facilitator, which may be a barrier for SMEs.</p> <p>Primarily focused on societal impact rather than technical fairness.</p> <p>Lacks comprehensive tools for technical fairness evaluation.</p>	<p>Conceptualising AI systems with societal and environmental considerations.</p> <p>Used for ethical evaluations of AI systems in public policy, education, and healthcare.</p> <p>Ensuring societal considerations are addressed in AI-based systems.</p>
Data Ethics Impact Assessment [144]	<p>Integrates data ethics into AI development processes.</p> <p>Focuses on the ethical implications of data usage and AI models.</p> <p>Provides an important framework for responsible AI development.</p>	<p>Limited to ethical assessments, not technical bias mitigation.</p> <p>May not address technical fairness challenges directly.</p> <p>Does not provide tools for bias correction or mitigation.</p>	<p>Ethical assessments in AI systems for public and private sectors.</p> <p>Used in ensuring data usage complies with ethical standards in sectors like healthcare and government.</p> <p>Ensuring ethical and fair data usage in AI systems for social justice.</p>
Veritas Fairness Assessment Methodology. [145]	<p>Developed for financial systems; focuses on fairness and transparency.</p> <p>Strong focus on transparency and accountability in financial applications.</p> <p>Tailored for regulatory and compliance environments.</p>	<p>Limited adoption outside the finance industry.</p> <p>Not widely applicable to non-financial systems.</p> <p>Limited support for non-financial use cases.</p>	<p>Fairness assessments in credit scoring and insurance systems.</p> <p>Used in fairness evaluations of automated financial decision-making systems.</p> <p>Ensuring fairness in automated decision-making systems in banking and finance.</p>
Assurance Cases for Fairness. [146]	<p>Provides structured arguments for fairness claims.</p> <p>Useful in establishing transparency and accountability for fairness claims.</p> <p>Focuses on providing evidence-based assurance for fairness.</p>	<p>Requires domain expertise and collaboration for effective implementation.</p> <p>May require significant resources to build effective assurance cases.</p> <p>Limited scalability to large AI systems.</p>	<p>Ensuring fairness in AI systems for healthcare and education.</p> <p>Used in verifying fairness in AI-based healthcare and educational systems.</p> <p>Ensuring trust and accountability in AI systems in sectors with high public scrutiny.</p>



Figure 9. AI fairness toolkits.

6.4. Robustness

Robustness refers to the capability of an algorithm or system to handle execution errors, unexpected inputs, or unfamiliar data effectively. It is a crucial factor influencing the dependability of AI systems in practical settings. Insufficient robustness can lead to unintended consequences or hazardous behaviour, compromising both safety and trust. Within the domain of machine learning, robustness covers various aspects. In this review, we categorise AI system vulnerabilities into three primary levels: data, algorithms, and system robustness.

1. **Data Level Robustness:** A model trained on limited datasets that do not reflect real-world variations may suffer significant performance degradation. One major challenge is a distributional shift, where the data seen during deployment differs from the training data, affecting model reliability [147]. This issue is particularly concerning in safety-critical domains. For example, in autonomous driving, AI models must function under a range of environmental conditions. While a system trained in sunny weather may perform well, its effectiveness in night time or rainy conditions could be severely reduced. To address this, researchers and industry professionals employ extensive testing and development strategies to improve AI perception under varying weather conditions, ensuring consistent performance [148,149].
2. **Algorithm-Level Robustness:** AI models can be vulnerable to adversarial attacks, where maliciously modified inputs deceive the system. These attacks have raised concerns in both academia and industry, leading to extensive research on threat classification and defence mechanisms [150–154]. Adversarial attacks can be categorised based on their timing:
 - **Decision-Time Attacks:** These involve modifying input samples in real-time to manipulate the model's predictions. Attackers may use such methods to bypass security mechanisms or impersonate legitimate users [155].
 - **Training-Time Attacks (Poisoning Attacks):** In this approach, adversaries introduce deceptive samples into the training data, influencing the model's learning process and altering its behaviour in specific situations [155].

Another important classification is based on the space in which attacks are conducted:

- **Feature-Space Attacks:** Traditional adversarial methods directly alter input features to deceive the model.
- **Problem-Space Attacks (Entity-Based Attacks):** Instead of modifying digital data, attackers alter physical objects to manipulate AI recognition. For example, a person wearing specially designed adversarial glasses could bypass a facial

recognition system [156,157]. Apart from adversarial attacks, model stealing (exploratory attacks) is another significant threat. These attacks do not directly alter model behaviour but extract knowledge about the AI system, which can later be exploited to craft more effective adversarial samples [158].

3. **System-Level Robustness:** AI systems must be designed to handle a wide range of unexpected or illegal inputs in real-world applications. Practical cases include the following:
 - **Unanticipated Inputs:** For instance, an image with an extremely high resolution might cause an AI-based image recognition system to crash.
 - **Sensor Interference:** In autonomous vehicles, a lidar system might misinterpret signals from other vehicles, leading to corrupted input data.
 - **Spoofing Attacks:** Attackers may use fake inputs—such as printed photos or masks—to deceive biometric authentication systems, raising security concerns [159]. To mitigate these risks, defensive mechanisms are categorised as either proactive or reactive [160]. Proactive defences aim to strengthen AI models against diverse inputs, making them inherently robust. Reactive defences focus on detecting adversarial samples or identifying anomalies in data distribution.
4. **Evaluating Robustness:** Assessing robustness is crucial for detecting vulnerabilities and managing risks. Two primary evaluation methods are robustness testing and mathematical verification.
 - **Robustness Testing** Testing plays a key role in validating AI robustness, just as it does in traditional software development. Techniques such as monkey testing—which uses randomised inputs to check system stability—can be applied to AI models [161]. Additionally, software testing methodologies have been adapted to assess AI resilience against adversarial attacks [162,163]. Another common method is performance testing (benchmarking), which evaluates model robustness using test datasets with varying distributions. One widely used metric is the minimal adversarial perturbation, which measures the smallest modification needed to mislead an AI model. Another key evaluation metric is the attack success rate, which reflects how easily an adversary can compromise the system [164,165].
 - **Mathematical Verification** Borrowed from formal verification methods, mathematical validation techniques are increasingly used to assess AI robustness. For instance, researchers derive certified lower bounds on the minimum distortion required for an adversarial attack—a measure of how resistant a model is to adversarial manipulations [166,167].

To enhance AI robustness, researchers and practitioners have developed specialised toolkits that help assess, mitigate, and defend against various vulnerabilities. These toolkits provide methods for robustness testing, adversarial attack detection, and model hardening, ensuring AI systems perform reliably across different conditions. Table 4 presents a comparative overview of key robustness toolkits, highlighting their functionalities, advantages, limitations, and practical use cases. Figure 10 further illustrates this comparison through a visual representation, providing additional clarity for evaluating these toolkits. The next section analyses how these trustworthiness factors are addressed across the entire AI pipeline, from data collection to model deployment. It formulates research questions and highlights practical challenges and best practices for ensuring trustworthy AI in real-world manufacturing settings.

Table 4. Comparison of AI robustness toolkits.

Toolkit	Pros	Cons	Use Cases
IBM ART [168]	Supports multiple ML frameworks Offers both attacks & defences Includes explainability tools	Can be complex to set up Some attacks are slow	Security testing for AI in finance & healthcare Adversarial training for robust AI models
CleverHans [169]	Well-documented Strong focus on adversarial attacks Open-source and widely used	Limited defence techniques Primarily TensorFlow-focused	Evaluating deep learning model security Research on adversarial robustness
Foolbox [170]	Simple API for adversarial attacks Works with PyTorch, TensorFlow, JAX Optimised for speed	Lacks built-in defence mechanisms Not as actively maintained as ART	Red teaming for AI security Benchmarking model vulnerability
MIT Robustness [171]	Designed for adversarial training PyTorch support Provides pre-trained robust models	Focused on image classification tasks Limited support for non-vision models	Adversarial training in computer vision Research in robustness techniques
DeepRobust [172]	Supports both graph and image-based AI models Covers attacks and defences Provides benchmark datasets	Requires deep understanding of adversarial learning Not as widely adopted as ART or CleverHans	AI security in social networks (graph AI) Robustness evaluation for medical AI models
AdverTorch [173]	PyTorch-based attack toolkit Various attack implementations Focus on adversarial training	Native PyTorch support Limited to PyTorch ecosystem	Developing adversarial defences in PyTorch Generating adversarial examples Robust training pipeline integration
Robustness Gym [174]	Robustness benchmarking suite Supports data transformations Model evaluation techniques	Modular and extensible Primarily NLP-focused	Evaluating NLP model robustness Stress-testing AI models in production Enhancing general AI model reliability
TRADES [175]	Trade-off between robustness and accuracy Adversarial training framework Defence against adversarial perturbations	Strong theoretical backing Requires deep ML expertise	Improving robustness in DNNs Research on robust AI training methods Defending against adversarial attacks
AutoAttack [176]	Ensemble adversarial attack method Automates attack selection Works on image classifiers	Strong attack performance Limited flexibility for defences	Benchmarking adversarial robustness Validating adversarial defences Testing robustness in computer vision
RobustBench [177]	Maintains a leaderboard of robust models Easy benchmarking of defences Open-source	Limited set of attacks compared to ART Mostly vision-focused	Benchmarking AI robustness in academia Comparing adversarial defences

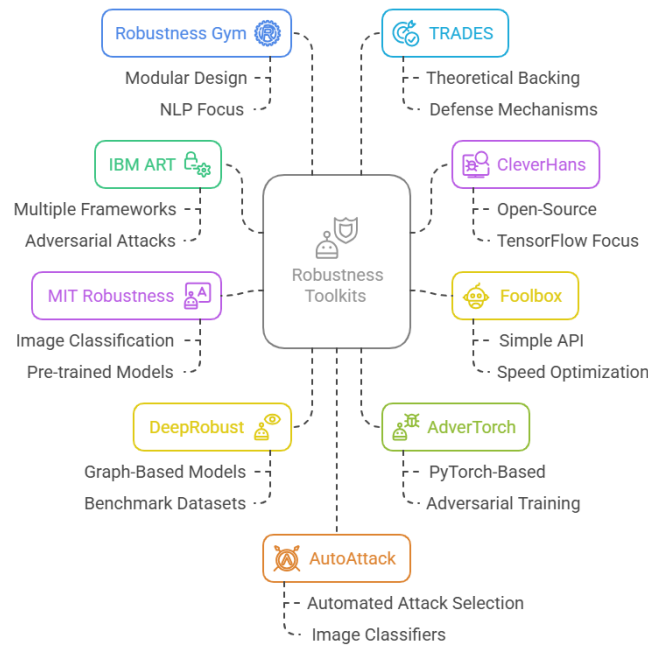


Figure 10. AI robustness toolkits.

7. Challenges in the AI Pipeline: From Data Collection to Model Deployment

While the principles of AI trustworthiness—transparency, fairness, robustness, and accountability—are well established in the literature, their practical realisation in manufacturing environments presents unique challenges and opportunities. This review bridges the theoretical and practical dimensions by mapping these core principles onto concrete factory-level scenarios. For example, transparency is operationalised through the deployment of explainable AI (XAI) tools that allow production engineers to interpret and validate machine learning predictions for quality control, thereby increasing trust in automated inspection systems. Fairness is addressed by monitoring and mitigating biases in predictive maintenance algorithms, ensuring that all equipment types and production lines receive equitable attention, rather than favouring those with more historical data. Robustness is exemplified by the implementation of adversarial testing protocols in digital twins, which simulate unexpected disruptions—such as sensor failures or supply chain shocks—to assess the resilience of AI-driven decision systems. Accountability is reinforced through the establishment of clear audit trails and responsibility matrices, enabling traceability of AI-driven decisions and facilitating compliance with regulatory standards. The practical realisation of these strategies often necessitates interdisciplinary collaboration. For instance, in the deployment of predictive maintenance systems, data scientists, manufacturing engineers, and ethicists have worked together to design algorithms that are not only technically robust but also transparent and fair in their recommendations. Such collaborations ensure that AI solutions are informed by domain expertise, ethical considerations, and operational realities, thereby enhancing both societal impact and user acceptance. By providing these scenario-based insights, this review demonstrates how the abstract dimensions of AI trustworthiness can be translated into actionable strategies and best practices for factory operations. This theoretical–practical integration not only supports the adoption of trustworthy AI in manufacturing but also aligns with the broader objectives of Industry 5.0, which emphasise human-centricity, sustainability, and ethical responsibility in industrial innovation. However, even with these strategies in place, manufacturing is undergoing a radical transformation driven by AI, particularly within the framework of Industry 5.0, which emphasises human–machine collaboration. AI technologies are central

to this transformation, enabling smarter, more efficient, and adaptable operations [1]. Yet, the adoption of AI is not without challenges. Each step in the AI pipeline—from data collection to model deployment and post-deployment monitoring—presents unique hurdles that impact the overall trustworthiness and effectiveness of AI solutions. Addressing these challenges requires a careful balance between technical innovation, robust ethical frameworks, and organisational transformation [178].

The development of AI models follows a structured pipeline shown in Figure 11, ensuring a systematic approach to creating reliable and trustworthy systems. The process begins with data collection, where data is gathered from various sources such as servers and IoT devices. This data forms the foundation for building AI models. Once collected, the data is securely stored in data storage systems, including databases like MySQL, PostgreSQL, and MongoDB, as well as cloud-based storage buckets such as Amazon S3, Google Cloud Storage, and Microsoft Azure Blob Storage. These systems ensure the data is organised, accessible, and ready for further processing [179].

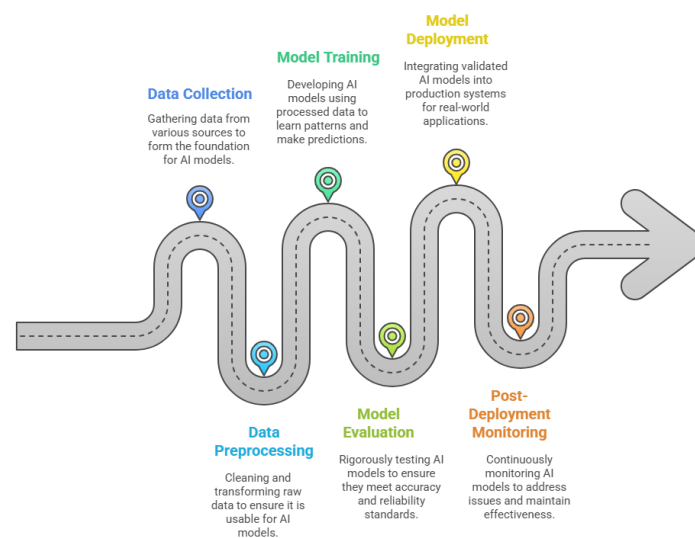


Figure 11. AI model development stages.

In the data processing stage, the raw data is filtered to remove noise and irrelevant information, and it is transformed into a usable format. This step ensures the data is clean, consistent, and ready for analysis. The processed data is then stored again in storage systems to maintain its integrity and accessibility for subsequent stages [180].

The next step is model training, where analysts and machine learning engineers use the processed data to train AI models. This involves designing algorithms that learn patterns and make predictions based on the data. For example, in manufacturing, AI models may be trained to predict equipment failures, optimise production schedules, or improve quality control. After training, the models undergo model evaluation, where their performance is rigorously tested to ensure they meet the required standards of accuracy, fairness, and robustness. This step is critical in manufacturing, as any inaccuracies in predictions or decisions can lead to costly disruptions or defects [181].

Finally, the validated models are integrated into production systems during the model deployment stage. Here, the models are used to perform real-world tasks, such as predictive maintenance, quality inspection, or supply chain optimisation. Post-deployment monitoring ensures that the models continue to perform effectively, addressing issues such as data drift, performance degradation, and cybersecurity risks. For instance, in manufacturing, continuous monitoring can help detect changes in production conditions or equipment behaviour that may affect the model's accuracy [182].

This structured pipeline ensures a systematic approach to AI development, enabling the creation of transparent and reliable AI systems in manufacturing. However, each stage of this pipeline presents unique challenges, such as ensuring data quality, addressing biases, and maintaining model robustness in dynamic manufacturing environments. These challenges must be addressed to fully realise the potential of AI in manufacturing and to build systems that are not only effective but also trustworthy and aligned with the principles of Industry 5.0. As AI systems progress through each stage of the manufacturing pipeline, trustworthiness can gradually erode—a process known as “trust leakage.” Small issues like data bias or reduced robustness, if not addressed early, may be amplified in later stages [27]. In the following subsections, we discuss how trust leakage can arise at each phase and strategies to mitigate these risks.

7.1. Data Collection

The data collection stage is fundamental to manufacturing AI, with data sourced from IoT sensors, legacy equipment, and digital systems. However, manufacturing data is often noisy, incomplete, and inconsistent, making it especially vulnerable to bias and privacy issues. If these challenges are not addressed early, they can propagate through the pipeline and compromise the fairness and reliability of AI models [183]. See Box 1.

Box 1. Illustrative example 1.

A global leader in sustainable manufacturing shown in Figure 12 has implemented an AI-driven smart factory to produce environmentally friendly products. The factory integrates AI across its operations, including supply chain management, production optimisation, predictive maintenance, and employee monitoring, with the goal of improving efficiency, reducing costs, and meeting sustainability targets. However, as the factory scales its AI systems, several challenges emerge. The organisation collects data from multiple sources, such as supplier databases, IoT sensors, and customer feedback, but faces issues with data interoperability due to inconsistent formats and schemas. For example, supplier data may use different terminologies, and IoT sensors from various vendors often produce incompatible data, leading to errors in supplier evaluation and production scheduling. Additionally, the reliability of external data sources becomes a concern when inaccurate information, such as incorrect carbon footprint data for materials, damages the organisation’s reputation when the error is discovered. Bias in historical data further complicates matters, as supplier selection models may favour long-term suppliers over new, innovative ones, and shopfloor automation systems might fail to recognise diverse accents and voices due to insufficiently inclusive training data. Predictive maintenance systems can also exhibit regional bias, relying on data from older machines predominantly used in one region, which results in inaccurate predictions for newer machines. Moreover, noisy or incomplete data from IoT sensors can cause false alarms in maintenance systems, leading to unnecessary production delays. Ethical concerns arise as the AI system monitors employee movements to ensure safety compliance but also flags workers for “low performance” based on arbitrary metrics, disproportionately affecting certain groups and raising privacy concerns. Finally, the AI system for demand forecasting struggles with concept drift, failing to adapt to sudden shifts in consumer preferences, which results in stockouts and lost sales.

The above example highlights the challenges, such as data bias, interoperability, and privacy, that can arise during the data collection steps. Therefore, this research raises the following questions related to this step:

- RQ1: How can data producers and owners implement interoperable data schemas to ensure data integrity?
- RQ2: What mechanisms best facilitate the extraction of unbiased, informative datasets from complex environments?
- RQ3: What types of biases are present in manufacturing datasets and data collection processes, and what strategies can be used to detect and address them while maintaining optimal performance?

- RQ4: What methods can be employed to gather unbiased and informative datasets from shop floor environments where human involvement is significant?
- RQ5: How can workers with limited AI expertise effectively evaluate algorithms for bias and fairness?

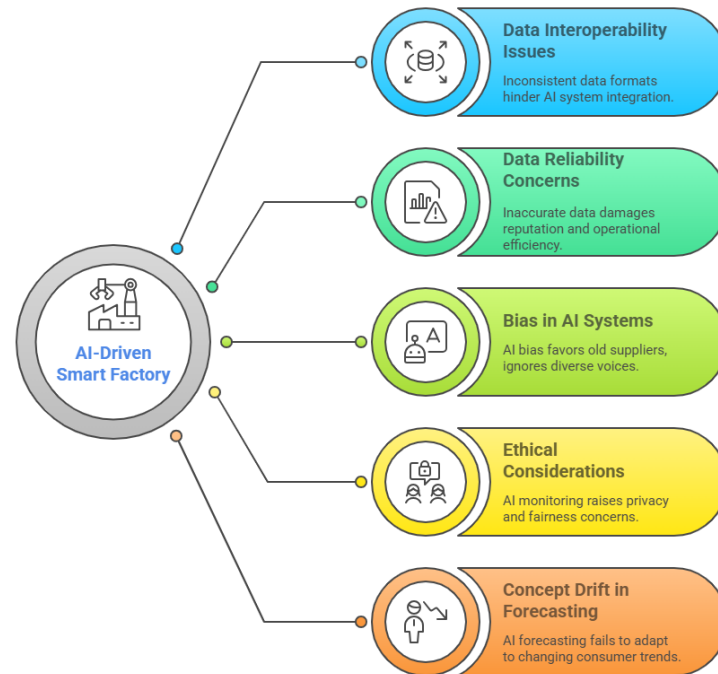


Figure 12. Navigating AI challenges in smart manufacturing.

7.2. Data Preprocessing

Data augmentation and preprocessing are essential for preparing manufacturing data—often sourced from IoT sensors, machine logs, and manual entries—for AI models. These steps clean, balance, and transform raw data to ensure reliability and accuracy. However, aggressive cleaning or augmentation can unintentionally introduce or amplify bias, impacting the fairness and robustness of downstream models [184]. See Box 2.

Box 2. Illustrative example 2.

Imagine a manufacturing company that produces automotive parts and wants to implement an AI system to predict product quality based on production parameters. The company collects data from various sources, including IoT sensors on machines, manual quality checks, and supplier records. However, several challenges arise shown in Figure 13 during data augmentation and preprocessing:

- **Labelling Issues:** The company needs labelled data to train its AI model, but quality labels are subjective and depend on the expertise of the quality inspectors. For instance, one inspector might classify a product as “acceptable”, while another might label it as “defective” under similar conditions. This inconsistency leads to uncertain labels, which can affect the model’s accuracy.
- **Data Imbalance:** Most of the products meet quality standards, resulting in an imbalanced dataset where defective products are under-represented. This imbalance can cause the AI model to overlook rare but critical defects.
- **Data Cleaning:** The IoT sensors occasionally produce noisy or incomplete data due to hardware malfunctions or network issues. For example, temperature readings from a sensor might show sudden spikes that do not reflect actual conditions, leading to incorrect predictions.
- **Feature Engineering:** The production team hypothesises that the humidity level in the factory might influence product quality. However, this data is not directly available and needs to be derived from existing environmental data, requiring domain expertise to create meaningful features.

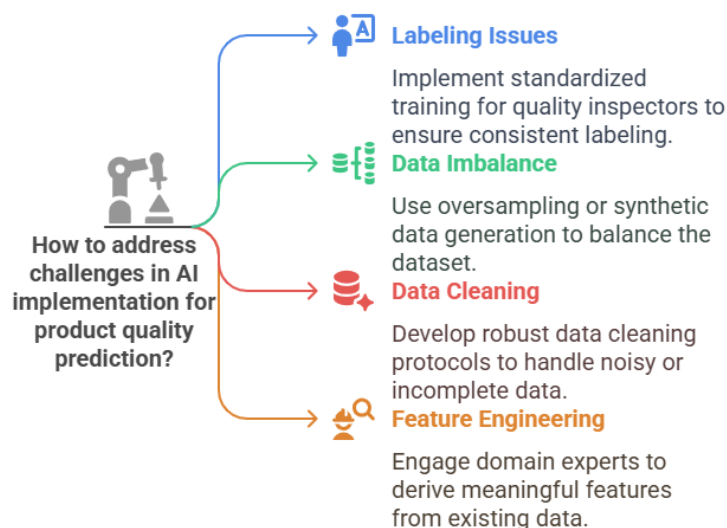


Figure 13. AI challenges in manufacturing.

To address the identified challenges in data preprocessing and augmentation for AI applications in manufacturing, the following research questions are proposed:

- RQ6: What methodologies can be developed to ensure consistent and accurate labelling of manufacturing data, particularly in scenarios where labels are subjective or context-dependent?
- RQ7: What are the most effective practices for addressing imbalanced datasets in manufacturing, and how can synthetic data generation techniques be optimised for such applications?
- RQ8: How can automated approaches be designed to detect and rectify noisy or incomplete data originating from IoT sensors and other manufacturing data sources?
- RQ9: What strategies can be employed to integrate domain knowledge into feature engineering processes while minimising the risk of introducing additional biases?
- RQ10: What frameworks or methodologies can be developed to identify and mitigate biases in manufacturing datasets, ensuring fair and unbiased AI-driven predictions?

7.3. Developing AI Models

Model development in manufacturing involves model selection, training, and deployment—each with unique challenges for trustworthiness. This stage is especially vulnerable to overfitting, lack of transparency, and bias amplification, particularly when performance is prioritised over interpretability. Addressing these risks is essential to ensure reliable and effective AI models.

1. Model selection is a crucial phase where the type of machine learning model is chosen, as it directly influences the model's interpretability and performance. Simpler models, such as decision trees, are often preferred in manufacturing due to their ease of understanding and transparency. However, this preference can sometimes lead to reduced accuracy, creating a trade-off between interpretability and performance. Additionally, computational limitations, particularly in resource-constrained environments, can restrict the use of more advanced models, further complicating the selection process [181].
2. Model training is another critical phase, where hyperparameters such as the depth of decision trees or the number of layers in a neural network are fine-tuned to enable the model to learn patterns effectively from the data. This phase is resource-intensive, with high computational and environmental costs, especially for large-scale models.

The process also requires skilled professionals to manage training effectively and avoid suboptimal results. While automated machine learning (AutoML) tools can simplify the training process, they often introduce challenges such as reduced transparency, potential bias, and overfitting, which can compromise the trustworthiness of the model [185].

3. Deployment phase focuses on integrating the model into real-world operations and ensuring its continued relevance. A significant challenge in this phase is addressing concept drift, where changes in the data distribution over time can render the model's predictions inaccurate. To mitigate this, organisations must implement mechanisms to detect when updates are needed and determine whether incremental updates or full retraining is more appropriate. This phase requires careful monitoring and adaptation to ensure the model remains effective in dynamic manufacturing environments [186]. See Box 3.

Box 3. Illustrative example 3.

A large automotive manufacturing company developed an AI model to predict machine failures on its assembly line, aiming to reduce unplanned downtime and optimise maintenance schedules. Initially, the model performed exceptionally well, using data from sensors monitoring machine vibrations, temperature, and operational cycles. The predictions allowed the company to schedule maintenance proactively, significantly reducing production delays. However, after several months, the model's accuracy began to decline, leading to unexpected machine breakdowns. Upon investigation, the team discovered that the supplier of a critical machine component had changed, resulting in slight variations in material properties. These changes in sensor readings were outside the scope of the model's training. Additionally, the production line was reconfigured to accommodate a new vehicle model, which altered the operational patterns of the machines, as shown in Figure 14.

To address the challenges highlighted above, this paper raises several important research questions.

- RQ11: How can manufacturing practitioners balance the trade-off between model interpretability and performance?
- RQ12: What are the most effective methods for detecting and managing concept drift in industrial applications?
- RQ13: How can organisations account for the environmental and financial costs of training AI models?

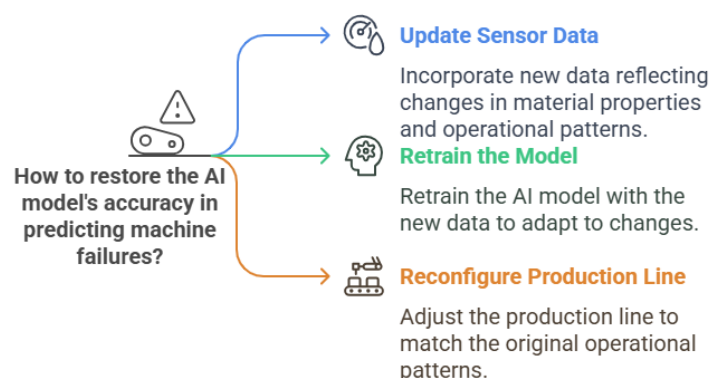


Figure 14. AI challenges in manufacturing.

8. Conclusions and Future Directions

AI is fundamentally transforming the manufacturing sector by enabling smarter, more efficient, and adaptive processes. As manufacturers increasingly depend on AI for decision-making, the importance of trustworthiness in these systems becomes paramount. Trustworthy AI is not only about technical accuracy but also about ensuring transparency,

fairness, robustness, and accountability. These principles are essential to prevent unintended consequences, such as biased decisions, lack of explainability, or system failures that could disrupt operations and erode stakeholder trust.

This paper has provided a comprehensive review of the current landscape of AI trustworthiness in manufacturing. It has highlighted the progress made in developing frameworks and tools for explainability, bias mitigation, and robust model deployment. However, it is clear that manufacturing environments are highly dynamic and complex, and no single solution can address all challenges. The effective implementation of trustworthy AI requires ongoing attention to ethical considerations, human values, and close collaboration between engineers, domain experts, and end-users. Only through such a holistic approach can the manufacturing industry fully harness the benefits of AI while minimising risks and ensuring responsible innovation.

Looking forward, the journey toward trustworthy AI in manufacturing must continue to evolve. Continuous monitoring and adaptation of AI models will be necessary to maintain their accuracy and fairness as manufacturing data and environments change over time. There is also a growing need to develop and adopt frameworks that integrate ethical, environmental, and financial considerations into every stage of the AI lifecycle. This integration will help ensure that AI systems not only perform well but also align with broader societal and sustainability goals.

Interdisciplinary collaboration will play a crucial role in this evolution. By bringing together AI specialists, manufacturing professionals, and ethicists, the industry can develop solutions that are both technically robust and ethically sound. Furthermore, regulatory compliance with emerging standards, such as the EU AI Act and ISO guidelines, will be essential for responsible AI deployment and for building stakeholder confidence.

A significant next phase for this research will involve the practical testing and evaluation of leading AI trustworthiness toolkits within real manufacturing scenarios. By rigorously assessing these toolkits, this research aims to identify which solutions are most effective and user-friendly in practice. The insights gained from this phase will provide actionable recommendations for industry adoption and will help bridge the gap between theoretical frameworks and real-world application.

While this study provides a comprehensive overview of AI trustworthiness in manufacturing, several limitations should be acknowledged. Methodologically, the reliance on literature review and illustrative, hypothetical case studies may limit the generalisability of the findings, as real-world complexities and sector-specific nuances may not be fully captured. Conceptually, the operationalisation of AI trustworthiness, though grounded in established frameworks, may be subject to interpretation and may not encompass all emerging dimensions as the field evolves. Contextually, the focus on manufacturing means that insights may not directly translate to other sectors or geographic regions with different regulatory, cultural, or operational landscapes. These limitations highlight the need for further empirical validation and cross-sectoral analysis.

To advance the field, future research should focus on developing robust, quantitative methods for assessing and benchmarking AI trustworthiness in manufacturing environments. This includes the creation of standardised metrics and tools for trust quantification, enabling objective comparison across different AI systems and deployment contexts. Comparative studies of AI regulatory frameworks across sectors and regions are also needed to identify best practices and inform the development of harmonised, context-sensitive standards. Additionally, longitudinal research tracking the long-term impacts of AI adoption on workforce dynamics, ethical outcomes, and organisational performance will provide valuable insights for both policymakers and industry leaders. Finally, there is a need for in-depth, real-world pilot studies that evaluate the effectiveness of trustworthiness tool-

its and frameworks in diverse manufacturing settings, thereby closing the gap between theoretical advances and practical implementation.

Author Contributions: Conceptualisation, M.N.A. and Z.A.F.; Methodology, M.N.A. and Z.A.F.; Writing—original draft preparation, M.N.A.; Writing—review and editing, M.N.A. and Z.A.F.; and Supervision, A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Innovate UK under High-Value Manufacturing Catapult (HVMC) project.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
GDPR	General Data Protection Regulation
AIX360	AI Explainability 360
SHAP	SHapley Additive exPlanations
LIME	Local Interpretable Model-agnostic Explanations
AIF360	AI Fairness 360
IBM ART	IBM Adversarial Robustness Toolbox
NIST	National Institute of Standards and Technology
OECD	Organisation for Economic Co-operation and Development
ISO	International Organization for Standardization
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
ALTAI	Assessment List for Trustworthy Artificial Intelligence
EU	European Union
IoT	Internet of Things
ML	Machine Learning
XAITK	Explainable AI Toolkit
HVMC	High-Value Manufacturing Catapult
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
AutoML	Automated Machine Learning
NLP	Natural Language Processing
DNN	Deep Neural Network
CNN	Convolutional Neural Network
RMF	Risk Management Framework

References

- Lu, Y. Industry 4.0: A survey on technologies, applications and open research issues. *J. Ind. Inf. Integr.* **2017**, *6*, 1–10. [[CrossRef](#)]
- Xu, L.D.; Duan, L. Big data for cyber physical systems in industry 4.0: A survey. *Enterp. Inf. Syst.* **2019**, *13*, 148–169. [[CrossRef](#)]
- Olsen, T.L.; Tomlin, B. Industry 4.0: Opportunities and challenges for operations management. *Manuf. Serv. Oper. Manag.* **2020**, *22*, 113–122. [[CrossRef](#)]
- Russell, S.J.; Norvig, P. *Artificial Intelligence: A Modern Approach*; Pearson: London, UK, 2016.
- Mitchell, T.M. *Machine Learning*; McGraw-Hill: New York, NY, USA, 1997.
- Lee, J.; Bagheri, B.; Kao, H.A. A cyber-physical systems architecture for industry 4.0-based manufacturing systems. *Manuf. Lett.* **2015**, *3*, 18–23. [[CrossRef](#)]

7. Xu, X.; Lu, Y.; Vogel-Heuser, B.; Wang, L. Industry 4.0 and Industry 5.0—Inception, conception and perception. *J. Manuf. Syst.* **2021**, *61*, 530–535. [CrossRef]
8. Ullah, A.S. Fundamental Issues of Concept Mapping Relevant to Discipline-Based Education: A Perspective of Manufacturing Engineering. *Educ. Sci.* **2019**, *9*, 228. [CrossRef]
9. Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv* **2017**, arXiv:1702.08608.
10. Hassija, V.; Chamola, V.; Mahapatra, A.; Singal, A.; Goel, D.; Huang, K.; Scardapane, S.; Spinelli, I.; Mahmud, M.; Hussain, A. Interpreting black-box models: A review on explainable artificial intelligence. *Cogn. Comput.* **2024**, *16*, 45–74. [CrossRef]
11. Greenberg, M.R. Energy policy and research: The underappreciation of trust. *Energy Res. Soc. Sci.* **2014**, *1*, 152–160. [CrossRef]
12. Floridi, L.; Cows, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. AI4People—An ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Mach.* **2018**, *28*, 689–707. [CrossRef]
13. Veale, M. A critical take on the policy recommendations of the EU high-level expert group on artificial intelligence. *Eur. J. Risk Regul.* **2020**, *11*, e1. [CrossRef]
14. AI, N. Artificial Intelligence Risk Management Framework (AI RMF 1.0). 2023. pp. 1–42. Available online: <https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf> (accessed on 7 July 2025).
15. Jobin, A.; Ienca, M.; Vayena, E. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* **2019**, *1*, 389–399. [CrossRef]
16. Butt, J. A Strategic Roadmap for the Manufacturing Industry to Implement Industry 4.0. *Designs* **2020**, *4*, 11. [CrossRef]
17. European Commission: DG Research and Innovation; Dixon-Declève, S.; Balland, P.-A.; Bria, F.; Charveriat, C.; Dunlop, K.; Giovannini, E.; Tataj, D.; Hidalgo, C.; Huang, A.; et al. *Industry 5.0, A Transformative Vision for Europe—Governing Systemic Transformations Towards a Sustainable Industry*; Publications Office of the European Union: Luxembourg, 2021.
18. Misra, N.N.; Dixit, Y.; Al-Mallahi, A.; Bhullar, M.S.; Upadhyay, R.; Martynenko, A. IoT, Big Data, and Artificial Intelligence in Agriculture and Food Industry. *IEEE Internet Things J.* **2022**, *9*, 6305–6324. [CrossRef]
19. Zheng, T.; Marco Ardolino, A.B.; Perona, M. The applications of Industry 4.0 technologies in manufacturing context: A systematic literature review. *Int. J. Prod. Res.* **2021**, *59*, 1922–1954. [CrossRef]
20. Angelopoulos, A.; Michailidis, E.T.; Nomikos, N.; Trakadas, P.; Hatziefremidis, A.; Voliotis, S.; Zahariadis, T. Tackling Faults in the Industry 4.0 Era—A Survey of Machine-Learning Solutions and Key Aspects. *Sensors* **2020**, *20*, 109. [CrossRef]
21. European Commission: DG Research and Innovation; Müller, J. *Enabling Technologies for Industry 5.0—Results of a Workshop with Europe’s Technology Leaders*; Publications Office: Luxembourg, 2020.
22. European Commission. Skills Agenda for Europe: Promoting Skills and Talent in the European Union. 2020. Available online: https://research-and-innovation.ec.europa.eu/knowledge-publications-tools-and-data/publications/all-publications/enabling-technologies-industry-50_en (accessed on 7 July 2025).
23. Claeys, G.; Tagliapietra, S.; Zachmann, G. *How to Make the European Green Deal Work*; JSTOR: New York, NY, USA, 2019; Volume 13.
24. Puaschunder, J.M. The legal and international situation of AI, robotics and big data with attention to healthcare. In *Report on Behalf of the European Parliament European Liberal Forum*; European Liberal Forum: Brussels, Belgium, 2019.
25. Marcus, J.S.; Martens, B.; Carugati, C.; Bucher, A.; Godlovitch, I. The European Health Data Space. IPOL | Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament Policy Department Studies. 2022. Available online: <https://ssrn.com/abstract=4300393> (accessed on 7 July 2025).
26. Nikolinakos, N.T. A European approach to excellence and trust: the 2020 white paper on artificial intelligence. In *EU Policy and Legal Framework for Artificial Intelligence, Robotics and Related Technologies—The AI Act*; Springer: Cham, Switzerland, 2023; pp. 211–280.
27. Ai, H. High-level expert group on artificial intelligence. *Ethics Guidel. Trust. AI* **2019**, *6*.
28. Mesarčík, M.; Solarova, S.; Podroužek, J.; Bieliková, M. Stance on the proposal for a regulation laying down harmonised rules on artificial intelligence—artificial intelligence act. In *OSF Preprints*; Center for Open Science: Charlottesville, VA, USA, 2022.
29. Cannarsa, M. Ethics guidelines for trustworthy AI. In *The Cambridge Handbook of Lawyering in the Digital Age*; Cambridge University Press: Cambridge, UK, 2021; pp. 283–297.
30. Ryan, M. The social and ethical impacts of artificial intelligence in agriculture: Mapping the agricultural AI literature. *AI Soc.* **2023**, *38*, 2473–2485. [CrossRef]
31. Wang, J.; Chuqiao, X.; Zhang, J.; Zhong, R. Big data analytics for intelligent manufacturing systems: A review. *J. Manuf. Syst.* **2021**, *62*, 738–752. [CrossRef]
32. Nahavandi, S. Industry 5.0—A human-centric solution. *Sustainability* **2019**, *11*, 4371. [CrossRef]
33. Rikap, C. Same End by Different Means: Google, Amazon, Microsoft and Facebook’s Strategies to Dominate Artificial Intelligence. 2023. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4472222 (accessed on 26 February 2025)
34. Habib ur Rehman, M.; Dirir, A.; Salah, K.; Damiani, E.; Svetinovic, D. TrustFed: A Framework for Fair and Trustworthy Cross-Device Federated Learning in IIoT. *IEEE Trans. Ind. Inform.* **2021**, 8485–8494. [CrossRef]

35. Vyhmeister, E.; Castane, G.; Östberg, P.O.; Thevenin, S. A responsible AI framework: Pipeline contextualisation. *AI Ethics* **2022**, *3*, 175–197. [CrossRef]
36. Vyhmeister, E.; Gonzalez-Castane, G.; Östberg, P.O. Risk as a driver for AI framework development on manufacturing. *AI Ethics* **2022**, *3*, 155–174. [CrossRef]
37. Wang, Q.; Liu, D.; Carmichael, M.; Aldini, S.; Lin, C.T. Computational Model of Robot Trust in Human Co-Worker for Physical Human-Robot Collaboration. *IEEE Robot. Autom. Lett.* **2022**, *7*, 3146–3153. [CrossRef]
38. Wu, J.; Drew, S.; Dong, F.; Zhu, Z.; Zhou, J. Topology-aware Federated Learning in Edge Computing: A Comprehensive Survey. *arXiv* **2023**. [CrossRef]
39. Reddy, P.; Pham, V.; B, P.; Deepa, N.; Dev, K.; Gadekallu, T.; Ruby, R.; Liyanage, M. Industry 5.0: A Survey on Enabling Technologies and Potential Applications. *J. Ind. Inf. Integr.* **2021**, *26*, 100257. [CrossRef]
40. Breque, M.; De Nul, L.; Petridis, A.; Directorate-General for Research and Innovation (European Commission). *Industry 5.0: Towards a Sustainable, Human-Centric and Resilient European Industry*; Publications Office of the European Union: Luxembourg, 2021; ISBN 978-92-76-25308-2. Available online: <https://op.europa.eu/en/publication-detail/-/publication/468a892a-5097-11eb-b59f-01aa75ed71a1/> (accessed on 7 July 2025).
41. Fahle, S.; Prinz, C.; Kuhlenkötter, B. Systematic review on machine learning (ML) methods for manufacturing processes—Identifying artificial intelligence (AI) methods for field application. *Procedia CIRP* **2020**, *93*, 413–418. [CrossRef]
42. IBM. How Is AI Being Used in Manufacturing? AI in Manufacturing. Available online: <https://www.ibm.com/think/topics/ai-in-manufacturing> (accessed on 7 July 2025).
43. Zonta, T.; da Costa, C.A.; da Rosa Righi, R.; de Lima, M.J.; da Trindade, E.S.; Li, G.P. Predictive maintenance in the Industry 4.0: A systematic literature review. *Comput. Ind. Eng.* **2020**, *150*, 106889. [CrossRef]
44. Qarout, Y.; Begg, M.; Fearon, L.; Russell, D.; Pietrow, N.; Rahman, M.; McLeod, S.; Chakravorti, N.; Winter, T.; Fortune, J.; et al. Trustworthy AI Framework: A Comprehensive Review of AI Standards Policies and a Practical Guideline to Their Application in Manufacturing. 2024. Available online: <https://www.the-mtc.org/api/sites/default/files/2024-07/Trustworthy%20AI%20Framework.pdf> (accessed on 26 February 2025).
45. European Commission; High-Level Expert Group on AI. Ethics Guidelines for Trustworthy AI. 2019. Available online: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> (accessed on 29 January 2025).
46. Reinhardt, K. Trust and trustworthiness in AI ethics. *AI Ethics* **2023**, *3*, 735–744. [CrossRef]
47. Idamia, S.; Benseddik, H. Advancing Industry 5.0: An Extensive Review of AI Integration. In *Industry 5.0 and Emerging Technologies: Transformation Through Technology and Innovations*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 3–21.
48. Dimitrakopoulos, G.; Varga, P.; Gutt, T.; Schneider, G.; Ehm, H.; Hoess, A.; Tauber, M.; Karathanasopoulou, K.; Lackner, A.; Delsing, J. Industry 5.0: Research Areas and Challenges With Artificial Intelligence and Human Acceptance. *IEEE Ind. Electron. Mag.* **2024**, *18*, 43–54. [CrossRef]
49. Keller, N. Cybersecurity Framework. 2013. Available online: <https://www.nist.gov/cyberframework> (accessed on 7 July 2025).
50. International Organization for Standardization. *ISO/IEC 27001:2022; Information Security, Cybersecurity and Privacy Protection—Information Security Management Systems—Requirements*. ISO: Geneva, Switzerland, 2022. Available online: <https://www.iso.org/standard/27001> (accessed on 16 June 2025).
51. Wachter, S.; Mittelstadt, B.; Floridi, L. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *Int. Data Priv. Law* **2017**, *7*, 76–99. [CrossRef]
52. Ala-Pietilä, P.; Bonnet, Y.; Bergmann, U.; Bielikova, M.; Bonefeld-Dahl, C.; Bauer, W.; Bouarfa, L.; Chatila, R.; Coeckelbergh, M.; Dignum, V.; et al. *The Assessment List for Trustworthy Artificial Intelligence (ALTAI)*; European Commission: Brussels, Belgium, 2020.
53. Martini, B.; Bellisario, D.; Coletti, P. Human-Centered and Sustainable Artificial Intelligence in Industry 5.0: Challenges and Perspectives. *Sustainability* **2024**, *16*, 5448. [CrossRef]
54. Yusuf, S.O.; Durodola, R.L.; Ocran, G.; Abubakar, J.E.; Echere, A.Z.; Paul-Adeleye, A.H. Challenges and opportunities in AI and digital transformation for SMEs: A cross-continental perspective. *World J. Adv. Res. Rev.* **2024**, *23*, 668–678. [CrossRef]
55. Wilson, H.J.; Daugherty, P.R. Collaborative intelligence: Humans and AI are joining forces. *Harv. Bus. Rev.* **2018**, *96*, 114–123.
56. Thomas, M. Dangerous Risks of Artificial Intelligence. Volume 6. Available online: <https://builtin.com/artificial-intelligence/risks-of-artificial-intelligence> (accessed on 1 March 2021)
57. Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. Machine bias. In *Ethics of Data and Analytics*; Auerbach Publications: Boca Raton, FL, USA, 2022; pp. 254–264.
58. Zhang, M. Google Photos Tags Two African-Americans as Gorillas Through Facial Recognition Software. 2015. Available online: <https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/> (accessed on 7 July 2025).
59. Dastin, J. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of Data and Analytics*; Auerbach Publications: Boca Raton, FL, USA, 2022; pp. 296–299.

60. Kohli, P.; Chadha, A. Enabling pedestrian safety using computer vision techniques: A case study of the 2018 uber inc. self-driving car crash. In Proceedings of the Future of Information and Communication Conference, San Francisco, CA, USA, 14–15 March 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 261–279.
61. Fan, W.; Liu, J.; Zhu, S.; Pardalos, P.M. Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (AIMDSS). *Ann. Oper. Res.* **2020**, *294*, 567–592. [[CrossRef](#)]
62. Elliott, A. *The Culture of AI: Everyday Life and the Digital Revolution*; Routledge: London, UK, 2019.
63. Marcus, G.; Davis, E. *Rebooting AI: Building Artificial Intelligence We Can Trust*; Vintage: London, UK, 2019.
64. Floridi, L.; Cowls, J. A unified framework of five principles for AI in society. In *Machine Learning and the City: Applications in Architecture and Urban Design*; Wiley Online Library: Hoboken, NJ, USA, 2022; pp. 535–545.
65. Floridi, L.; Cowls, J.; King, T.C.; Taddeo, M. How to design AI for social good: Seven essential factors. In *Ethics, Governance, and Policies in Artificial Intelligence*; Springer: Cham, Switzerland, 2021; pp. 125–151.
66. Pichai, S. AI at Google: Our principles. *Keyword* **2018**, *7*, 1–3.
67. UNI Global Union. Top 10 principles for ethical artificial intelligence. *The Future World of Work*, UNI Global Union: Nyon, Switzerland, 2017. Available online: <https://uniglobalunion.org/report/10-principles-for-ethical-artificial-intelligence/> (accessed on 7 July 2025).
68. Zeng, Y.; Lu, E.; Huangfu, C. Linking artificial intelligence principles. *arXiv* **2018**, arXiv:1812.04814.
69. Hagendorff, T. The ethics of AI ethics: An evaluation of guidelines. *Minds Mach.* **2020**, *30*, 99–120. [[CrossRef](#)]
70. Kaur, D.; Uslu, S.; Durrresi, A. Requirements for trustworthy artificial intelligence—a review. In Proceedings of the Advances in Networked-Based Information Systems: The 23rd International Conference on Network-Based Information Systems (NBIS-2020) 23, Victoria, BC, Canada, 31 August–2 September 2020; Springer: Berlin/Heidelberg, Germany, 2021; pp. 105–115.
71. Kumar, A.; Braud, T.; Tarkoma, S.; Hui, P. Trustworthy AI in the age of pervasive computing and big data. In Proceedings of the 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), Austin, TX, USA, 23–27 March 2020; pp. 1–6.
72. Smuha, N.A. The EU approach to ethics guidelines for trustworthy artificial intelligence. *Comput. Law Rev. Int.* **2019**, *20*, 97–106. [[CrossRef](#)]
73. Daugherty, P.R.; Wilson, H.J. *Human+ Machine: Reimagining Work in the Age of AI*; Harvard Business Press: Brighton, MA, USA, 2018.
74. Goodman, B.; Flaxman, S. European Union regulations on algorithmic decision-making and a “right to explanation”. *AI Mag.* **2017**, *38*, 50–57. [[CrossRef](#)]
75. Dignum, V. Responsible Artificial Intelligence: Designing AI for Human Values. *ITU J. ICT Discov.* **2018**, *1*, 1–8. Available online: https://www.itu.int/dms_pub/itu-s/opb/journal/S-JOURNAL-ICTF.VOL1-2018-1-P01-PDF-E.pdf (accessed on 5 March 2025).
76. Shin, D.; Park, Y.J. Role of fairness, accountability, and transparency in algorithmic affordance. *Comput. Hum. Behav.* **2019**, *98*, 277–284. [[CrossRef](#)]
77. Flores, A.W.; Bechtel, K.; Lowenkamp, C.T. False positives, false negatives, and false analyses: A rejoinder to machine bias: There’s software used across the country to predict future criminals. and it’s biased against blacks. *Fed. Probat.* **2016**, *80*, 38.
78. Wieringa, M. What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; pp. 1–18.
79. Islam, M.R.; Ahmed, M.U.; Barua, S.; Begum, S. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Appl. Sci.* **2022**, *12*, 1353. [[CrossRef](#)]
80. Chou, Y.L.; Moreira, C.; Bruza, P.; Ouyang, C.; Jorge, J. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Inf. Fusion* **2022**, *81*, 59–83. [[CrossRef](#)]
81. Ahmed, I.; Jeon, G.; Piccialli, F. From artificial intelligence to explainable artificial intelligence in industry 4.0: A survey on what, how, and where. *IEEE Trans. Ind. Inform.* **2022**, *18*, 5031–5042. [[CrossRef](#)]
82. Khosravi, H.; Shum, S.B.; Chen, G.; Conati, C.; Tsai, Y.S.; Kay, J.; Knight, S.; Martinez-Maldonado, R.; Sadiq, S.; Gašević, D. Explainable artificial intelligence in education. *Comput. Educ. Artif. Intell.* **2022**, *3*, 100074. [[CrossRef](#)]
83. Rawal, A.; McCoy, J.; Rawat, D.B.; Sadler, B.M.; Amant, R.S. Recent advances in trustworthy explainable artificial intelligence: Status, challenges, and perspectives. *IEEE Trans. Artif. Intell.* **2021**, *3*, 852–866. [[CrossRef](#)]
84. Albahri, A.S.; Duhaim, A.M.; Fadhel, M.A.; Alnoor, A.; Baqer, N.S.; Alzubaidi, L.; Albahri, O.S.; Alamoodi, A.H.; Bai, J.; Salhi, A.; et al. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Inf. Fusion* **2023**, *96*, 156–191. [[CrossRef](#)]
85. Albahri, A.; Al-Qaysi, Z.; Alzubaidi, L.; Alnoor, A.; Albahri, O.; Alamoodi, A.; Bakar, A.A. A Systematic Review of Using Deep Learning Technology in the Steady-State Visually Evoked Potential-Based Brain-Computer Interface Applications: Current Trends and Future Trust Methodology. *Int. J. Telemed. Appl.* **2023**, *2023*, 7741735. [[CrossRef](#)]

86. Velmurugan, M.; Ouyang, C.; Moreira, C.; Sindhgatta, R. Evaluating stability of post-hoc explanations for business process predictions. In Proceedings of the International Conference on Service-Oriented Computing, Dubai, United Arab Emirates, 22–25 November 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 49–64.
87. Selbst, A.; Powles, J. “Meaningful information” and the right to explanation. In Proceedings of the Conference on Fairness, Accountability and Transparency, New York, NY, USA, 23–24 February 2018; pp. 48–48.
88. Velmurugan, M.; Ouyang, C.; Moreira, C.; Sindhgatta, R. Evaluating fidelity of explainable methods for predictive process analytics. In Proceedings of the International Conference on Advanced Information Systems Engineering, Melbourne, VIC, Australia, 28 June–2 July 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 64–72.
89. Sreedharan, S.; Srivastava, S.; Kambhampati, S. Using state abstractions to compute personalized contrastive explanations for AI agent behavior. *Artif. Intell.* **2021**, *301*, 103570. [[CrossRef](#)]
90. Shin, D. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *Int. J. Hum.-Comput. Stud.* **2021**, *146*, 102551. [[CrossRef](#)]
91. Lim, W.X.; Chen, Z.; Ahmed, A. The adoption of deep learning interpretability techniques on diabetic retinopathy analysis: A review. *Med Biol. Eng. Comput.* **2022**, *60*, 633–642. [[CrossRef](#)] [[PubMed](#)]
92. Huang, Y.; Chen, D.; Zhao, W.; Lv, Y.; Wang, S. Deep patch learning algorithms with high interpretability for regression problems. *Int. J. Intell. Syst.* **2022**, *37*, 8239–8276. [[CrossRef](#)]
93. Yang, C.; Rangarajan, A.; Ranka, S. Global model interpretation via recursive partitioning. In Proceedings of the 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), Exeter, UK, 28–30 June 2018; pp. 1563–1570.
94. Moreira, C.; Chou, Y.L.; Velmurugan, M.; Ouyang, C.; Sindhgatta, R.; Bruza, P. LINDA-BN: An interpretable probabilistic approach for demystifying black-box predictive models. *Decis. Support Syst.* **2021**, *150*, 113561. [[CrossRef](#)]
95. Lyu, D.; Yang, F.; Kwon, H.; Dong, W.; Yilmaz, L.; Liu, B. Tdm: Trustworthy decision-making via interpretability enhancement. *IEEE Trans. Emerg. Top. Comput. Intell.* **2021**, *6*, 450–461. [[CrossRef](#)]
96. Reed, C. How should we regulate artificial intelligence? *Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci.* **2018**, *376*, 20170360. [[CrossRef](#)]
97. Wickramanayake, B.; He, Z.; Ouyang, C.; Moreira, C.; Xu, Y.; Sindhgatta, R. Building interpretable models for business process prediction using shared and specialised attention mechanisms. *Knowl.-Based Syst.* **2022**, *248*, 108773. [[CrossRef](#)]
98. Thiebes, S.; Lins, S.; Sunyaev, A. Trustworthy artificial intelligence. *Electron. Mark.* **2021**, *31*, 447–464. [[CrossRef](#)]
99. Sindhgatta, R.; Ouyang, C.; Moreira, C. Exploring interpretability for predictive process analytics. In Proceedings of the International Conference on Service-Oriented Computing, Dubai, United Arab Emirates, 14–17 December 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 439–447.
100. Bhatt, U.; Xiang, A.; Sharma, S.; Weller, A.; Taly, A.; Jia, Y.; Ghosh, J.; Puri, R.; Moura, J.M.; Eckersley, P. Explainable machine learning in deployment. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; pp. 648–657.
101. IBM. AI Explainability 360. 2025. Available online: <https://research.ibm.com/blog/ai-explainability-360> (accessed on 18 March 2025).
102. Ribeiro, M.T.; Singh, S.; Guestrin, C. Why should I trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, 2016; pp. 1135–1144. Available online: <https://dl.acm.org/doi/10.1145/2939672.2939778> (accessed on 7 July 2025).
103. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 4765–4774. Available online: <https://arxiv.org/abs/1705.07874> (accessed on 7 July 2025).
104. Chen, J.; Kanan, C.; Fowlkes, C.C. XAITK Saliency: An Open Source Explainable AI Toolkit for Visual Saliency. Available online: <https://ojs.aaai.org/index.php/AAAI/article/view/26871> (accessed on 18 March 2025).
105. Yin, H.; Xie, K.; Zhou, Y.; Liu, J.; Zhang, Z.; Liu, P.; Wang, W.; Zhao, Y. Fidelity in Explanation Methods: A Comprehensive Study. *arXiv* **2022**, arXiv:2202.06861.
106. Ieracitano, C.; Mammone, N.; Hussain, A.; Morabito, F.C. A novel explainable machine learning approach for EEG-based brain-computer interface systems. *Neural Comput. Appl.* **2022**, *34*, 11347–11360. [[CrossRef](#)]
107. Ras, G.; Xie, N.; Van Gerven, M.; Doran, D. Explainable deep learning: A field guide for the uninitiated. *J. Artif. Intell. Res.* **2022**, *73*, 329–396. [[CrossRef](#)]
108. De Waal, A.; Joubert, J.W. Explainable Bayesian networks applied to transport vulnerability. *Expert Syst. Appl.* **2022**, *209*. [[CrossRef](#)]
109. Mao, C.; Lin, R.; Towey, D.; Wang, W.; Chen, J.; He, Q. Trustworthiness prediction of cloud services based on selective neural network ensemble learning. *Expert Syst. Appl.* **2021**, *168*, 114390. [[CrossRef](#)]

110. Srinivasan, R.; González, B.S.M. The role of empathy for artificial intelligence accountability. *J. Responsible Technol.* **2022**, *9*, 100021. [[CrossRef](#)]
111. Choudhury, A.; Asan, O. Impact of accountability, training, and human factors on the use of artificial intelligence in healthcare: Exploring the perceptions of healthcare practitioners in the US. *Hum. Factors Healthc.* **2022**, *2*, 100021. [[CrossRef](#)]
112. Fong, S.; Dey, N.; Joshi, A. ICT analysis and applications. In Proceedings of the ICT4SD, Goa, India, 23–24 July 2020; Volume 2.
113. Cruz, B.S.; de Oliveira Dias, M. Crashed boeing 737-max: Fatalities or malpractice. *GSJ* **2020**, *8*, 2615–2624.
114. Schwartz, R.; Schwartz, R.; Vassilev, A.; Greene, K.; Perine, L.; Burt, A.; Hall, P. In *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*; US Department of Commerce, National Institute of Standards and Technology: Gaithersburg, MD, USA, 2022; Volume 3.
115. Gevaert, C.M.; Carman, M.; Rosman, B.; Georgiadou, Y.; Soden, R. Fairness and accountability of AI in disaster risk management: Opportunities and challenges. *Patterns* **2021**, *2*, 100363. [[CrossRef](#)]
116. Königstorfer, F.; Thalmann, S. AI Documentation: A path to accountability. *J. Responsible Technol.* **2022**, *11*, 100043. [[CrossRef](#)]
117. Rahwan, I.; Cebrian, M.; Obradovich, N.; Bongard, J.; Bonnefon, J.F.; Breazeal, C.; Crandall, J.W.; Christakis, N.A.; Couzin, I.D.; Jackson, M.O.; et al. Machine behaviour. *Nature* **2019**, *568*, 477–486. [[CrossRef](#)]
118. Morley, J.; Floridi, L.; Kinsey, L.; Elhalal, A. From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Sci. Eng. Ethics* **2020**, *26*, 2141–2168. [[CrossRef](#)] [[PubMed](#)]
119. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A survey of methods for explaining black box models. *ACM Comput. Surv. (CSUR)* **2018**, *51*, 1–42. [[CrossRef](#)]
120. Omeiza, D.; Web, H.; Jirotko, M.; Kunze, L. Towards accountability: Providing intelligible explanations in autonomous driving. In Proceedings of the 2021 IEEE Intelligent Vehicles Symposium (IV), Nagoya, Japan, 11–17 July 2021; pp. 231–237.
121. Kadambi, A. Achieving fairness in medical devices. *Science* **2021**, *372*, 30–31. [[CrossRef](#)] [[PubMed](#)]
122. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–35. [[CrossRef](#)]
123. Madaio, M.; Egede, L.; Subramonyam, H.; Wortman Vaughan, J.; Wallach, H. Assessing the fairness of ai systems: Ai practitioners' processes, challenges, and needs for support. In Proceedings of the ACM on Human-Computer Interaction, New Orleans, LA, USA, 29 April–5 May 2022; Volume 6, pp. 1–26.
124. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [[CrossRef](#)]
125. von Zahn, M.; Feuerriegel, S.; Kuehl, N. The cost of fairness in AI: Evidence from e-commerce. *Bus. Inf. Syst. Eng.* **2022**, *64*, 335–348. [[CrossRef](#)]
126. Pastaltzidis, I.; Dimitriou, N.; Quezada-Tavarez, K.; Aidinlis, S.; Marquenie, T.; Gurzawska, A.; Tzovaras, D. Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, 21–24 June 2022; pp. 2302–2314.
127. Chouldechova, A.; Benavides-Prado, D.; Fialko, O.; Vaithianathan, R. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In Proceedings of the Conference on Fairness, Accountability and Transparency, New York, NY, USA, 23–24 February 2018; pp. 134–148.
128. Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; Zemel, R. Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, Cambridge, MA, USA, 8–10 January 2012; pp. 214–226.
129. Oneto, L.; Chiappa, S. *Recent Trends in Learning from Data*; Springer: Berlin/Heidelberg, Germany, 2020.
130. Kusner, M.J.; Loftus, J.; Russell, C.; Silva, R. Counterfactual fairness. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.
131. Grari, V.; Lamprier, S.; Detyniecki, M. Adversarial learning for counterfactual fairness. *Mach. Learn.* **2023**, *112*, 741–763. [[CrossRef](#)]
132. Santos, F.P.; Santos, F.C.; Paiva, A.; Pacheco, J.M. Evolutionary dynamics of group fairness. *J. Theor. Biol.* **2015**, *378*, 96–102. [[CrossRef](#)]
133. Khalili, M.M.; Zhang, X.; Abroshan, M. Fair sequential selection using supervised learning models. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 28144–28155.
134. Zheng, Y.; Wang, S.; Zhao, J. Equality of opportunity in travel behavior prediction with deep neural networks and discrete choice models. *Transp. Res. Part C Emerg. Technol.* **2021**, *132*, 103410. [[CrossRef](#)]
135. Besse, P.; del Barrio, E.; Gordaliza, P.; Loubes, J.M.; Risser, L. A survey of bias in machine learning through the prism of statistical parity. *Am. Stat.* **2022**, *76*, 188–198. [[CrossRef](#)]
136. Feuerriegel, S.; Dolata, M.; Schwabe, G. Fair AI: Challenges and opportunities. *Bus. Inf. Syst. Eng.* **2020**, *62*, 379–384. [[CrossRef](#)]
137. Chen, Y.; Huerta, E.; Duarte, J.; Harris, P.; Katz, D.S.; Neubauer, M.S.; Diaz, D.; Mokhtar, F.; Kansal, R.; Park, S.E.; et al. A FAIR and AI-ready Higgs boson decay dataset. *Sci. Data* **2022**, *9*, 31. [[CrossRef](#)]
138. AI Fairness 360. AI Fairness 360. 2025. Available online: <https://research.ibm.com/blog/ai-fairness-360> (accessed on 7 July 2025).
139. Fairlearn. Fairlearn. 2025. Available online: <https://fairlearn.org/> (accessed on 13 March 2025).

140. What-If Tool. What-If Tool. 2025. Available online: <https://pair-code.github.io/what-if-tool/> (accessed on 7 July 2025).
141. Aequitas. Aequitas. 2025. Available online: <https://arxiv.org/abs/1811.05577> (accessed on 13 March 2025).
142. Google Research. ML Fairness Gym: A Tool for Exploring Long-Term Impacts of Machine Learning Systems. 2019. Available online: <https://research.google/blog/ml-fairness-gym-a-tool-for-exploring-long-term-impacts-of-machine-learning-systems/> (accessed on 7 July 2025).
143. AI Now Institute. Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability. 2022. Available online: <https://ainowinstitute.org/publications/algorithmic-impact-assessments-report-2> (accessed on 7 July 2025).
144. UK Government. Data Ethics Framework 2020. 2020. Available online: https://assets.publishing.service.gov.uk/media/5f74a4958fa8f5188dad0e99/Data_Ethics_Framework_2020.pdf (accessed on 7 July 2025).
145. Swiss Re. A Journey into Responsible AI. 2023. Available online: <https://www.swissre.com/institute/> (accessed on 7 July 2025).
146. Hauer, M.P.; Adler, R.; Zweig, K. Assuring fairness of algorithmic decision making. In Proceedings of the 2021 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), Porto de Galinhas, Brazil, 12–16 April 2021; pp. 110–113.
147. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. Concrete problems in AI safety. *arXiv* **2016**, arXiv:1606.06565.
148. Tan, X.; Xu, K.; Cao, Y.; Zhang, Y.; Ma, L.; Lau, R.W. Night-time scene parsing with a large real dataset. *IEEE Trans. Image Process.* **2021**, *30*, 9085–9098. [[CrossRef](#)]
149. Zhang, M.; Zhang, Y.; Zhang, L.; Liu, C.; Khurshid, S. Deeproad: Gan-based metamorphic testing and input validation framework for autonomous driving systems. In Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering, Montpellier, France, 3–7 September 2018; pp. 132–142.
150. Akhtar, N.; Mian, A. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access* **2018**, *6*, 14410–14430. [[CrossRef](#)]
151. Chakraborty, A.; Alam, M.; Dey, V.; Chattopadhyay, A.; Mukhopadhyay, D. Adversarial attacks and defences: A survey. *arXiv* **2018**, arXiv:1810.00069. [[CrossRef](#)]
152. Li, Y.; Jiang, Y.; Li, Z.; Xia, S.T. Backdoor learning: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *35*, 5–22. [[CrossRef](#)]
153. Silva, S.H.; Najafirad, P. Opportunities and challenges in deep learning adversarial robustness: A survey. *arXiv* **2020**, arXiv:2007.00753.
154. Yuan, X.; He, P.; Zhu, Q.; Li, X. Adversarial examples: Attacks and defenses for deep learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 2805–2824. [[CrossRef](#)] [[PubMed](#)]
155. Vorobeychik, Y.; Kantarcioglu, M. *Adversarial Machine Learning*; Morgan & Claypool Publishers: San Rafael, CA, USA, 2018.
156. Tong, L.; Li, B.; Hajaj, C.; Xiao, C.; Zhang, N.; Vorobeychik, Y. Improving robustness of {ML} classifiers against realizable evasion attacks using conserved features. In Proceedings of the 28th USENIX Security Symposium (USENIX Security 19), Santa Clara, CA, USA, 14–16 August 2019; pp. 285–302.
157. Wu, T.; Tong, L.; Vorobeychik, Y. Defending against physically realizable attacks on image classification. *arXiv* **2019**, arXiv:1909.09552.
158. Tramèr, F.; Zhang, F.; Juels, A.; Reiter, M.K.; Ristenpart, T. Stealing machine learning models via prediction {APIs}. In Proceedings of the 25th USENIX security symposium (USENIX Security 16), Austin, TX, USA, 10–12 August 2016; pp. 601–618.
159. Ramachandra, R.; Busch, C. Presentation attack detection methods for face recognition systems: A comprehensive survey. *ACM Comput. Surv. (CSUR)* **2017**, *50*, 1–37. [[CrossRef](#)]
160. Machado, G.R.; Silva, E.; Goldschmidt, R.R. Adversarial machine learning in image classification: A survey toward the defender’s perspective. *ACM Comput. Surv. (CSUR)* **2021**, *55*, 1–38. [[CrossRef](#)]
161. Exforsys. What Is Monkey Testing. 2011. Available online: <http://www.exforsys.com/tutorials/testing-types/monkey-testing.html> (accessed on 7 July 2025).
162. Ma, L.; Juefei-Xu, F.; Zhang, F.; Sun, J.; Xue, M.; Li, B.; Chen, C.; Su, T.; Li, L.; Liu, Y.; et al. Deepgauge: Multi-granularity testing criteria for deep learning systems. In Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering, Montpellier, France, 3–7 September 2018; pp. 120–131.
163. Pei, K.; Cao, Y.; Yang, J.; Jana, S. Deepxplore: Automated whitebox testing of deep learning systems. In Proceedings of the 26th Symposium on Operating Systems Principles, Shanghai, China, 28–31 October 2017; pp. 1–18.
164. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2017; pp. 39–57.
165. Su, D.; Zhang, H.; Chen, H.; Yi, J.; Chen, P.Y.; Gao, Y. Is robustness the cost of accuracy?—a comprehensive study on the robustness of 18 deep image classification models. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 631–648.

166. Boopathy, A.; Weng, T.W.; Chen, P.Y.; Liu, S.; Daniel, L. Cnn-cert: An efficient framework for certifying robustness of convolutional neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 3240–3247.
167. Zhang, H.; Weng, T.W.; Chen, P.Y.; Hsieh, C.J.; Daniel, L. Efficient neural network robustness certification with general activation functions. *Adv. Neural Inf. Process. Syst.* **2018**, *31*.
168. Nicolae, M.I.; Sinn, M.; Tran, M.N.; Buesser, B.; Rawat, A.; Wistuba, M.; Zantedeschi, V.; Baracaldo, N.; Chen, B.; Ludwig, H.; et al. Adversarial Robustness Toolbox v1.0.0. *arXiv* **2018**, arXiv:1807.01069.
169. Papernot, N.; Faghri, F.; Carlini, N.; Goodfellow, I.; Feinman, R.; Kurakin, A.; Xie, C.; Sharma, Y.; Brown, T.; Roy, A.; et al. Technical Report on the CleverHans v2.1.0 Adversarial Examples Library. *arXiv* **2018**, arXiv:1610.00768.
170. Rauber, J.; Brendel, W.; Bethge, M. Foolbox: A Python toolbox to benchmark the robustness of machine learning models. *arXiv* **2018**, arXiv:1707.04131.
171. Soklaski, R.; Goodwin, J.; Brown, O.; Yee, M.; Matterer, J. Tools and Practices for Responsible AI Engineering. *arXiv* **2022**, arXiv:2201.05647.
172. Li, Y.; Jin, W.; Xu, H.; Tang, J. DeepRobust: A PyTorch Library for Adversarial Attacks and Defenses. *arXiv* **2020**, arXiv:2005.06149.
173. Chen, J.; Su, Y.; Song, Z.; Wang, B.; Tang, R.; Zhao, S.; Yang, Y.; Zhang, J.; Kumar, M.; Xu, H. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv* **2019**, arXiv:1902.07623.
174. Radford, A.; Kim, J.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, K.; Clark, J.; Krueger, G.; Chen, M.; Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision. *arXiv* **2021**, arXiv:2101.04840.
175. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shinn, E.; Ibarz, J.; Wu, L.; et al. Language Models are Unsupervised Multitask Learners. *arXiv* **2019**, arXiv:1901.08573.
176. Liu, Z.; Lin, Y.; Cao, Y.; Zhang, F.; Zhang, Z.; Wang, X.; Zhang, T.; Chen, X.; Zeng, R.; Yu, L. Data-efficient image transformers for downstream tasks. *arXiv* **2022**, arXiv:2203.05154.
177. Yu, H.; Zheng, F.; Hu, H.; Gu, S.; Zhang, S.; Li, H. Learning to Learn with Conditional Class Dependencies. *arXiv* **2020**, arXiv:2010.09670.
178. He, M.; Li, Z.; Liu, C.; Shi, D.; Tan, Z. Deployment of Artificial Intelligence in Real-World Practice: Opportunity and Challenge. *Asia-Pac. J. Ophthalmol.* **2020**, *9*, 299–307. [[CrossRef](#)]
179. Peres, R.S.; Jia, X.; Lee, J.; Sun, K.; Colombo, A.W.; Barata, J. Industrial artificial intelligence in industry 4.0-systematic review, challenges and outlook. *IEEE Access* **2020**, *8*, 220121–220139. [[CrossRef](#)]
180. Ahsan, M.; Hon, S.T.; Albarbar, A. Development of Novel Big Data Analytics Framework for Smart Clothing. *IEEE Access* **2020**, *8*, 146376–146394. [[CrossRef](#)]
181. Wang, P.; Wang, K.; Wang, D.; Liu, H. The Impact of Manufacturing Transformation in Digital Economy Under Artificial Intelligence. *IEEE Access* **2024**, *12*, 63417–63424. [[CrossRef](#)]
182. Colombi, L.; Gilli, A.; Dahdal, S.; Boleac, I.; Tortonesi, M.; Stefanelli, C.; Vignoli, M. A Machine Learning Operations Platform for Streamlined Model Serving in Industry 5.0. In Proceedings of the NOMS 2024-2024 IEEE Network Operations and Management Symposium, Seoul, Republic of Korea, 6–10 May 2024; pp. 1–6. [[CrossRef](#)]
183. Yu, W.; Liu, Y.; Dillon, T.; Rahayu, W.; Mostafa, F. An Integrated Framework for Health State Monitoring in a Smart Factory Employing IoT and Big Data Techniques. *IEEE Internet Things J.* **2022**, *9*, 2443–2454. [[CrossRef](#)]
184. Hariharakrishnan, J.; Mohanavalli, S.; Srividya.; Sundhara Kumar, K.B. Survey of pre-processing techniques for mining big data. In Proceedings of the 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP), Chennai, India, 10–11 January 2017; pp. 1–5. [[CrossRef](#)]
185. Shahriar, S.; Allana, S.; Hazratifard, S.M.; Dara, R. A Survey of Privacy Risks and Mitigation Strategies in the Artificial Intelligence Life Cycle. *IEEE Access* **2023**, *11*, 61829–61854. [[CrossRef](#)]
186. Kemnitz, J.; Weissenfeld, A.; Schoeffl, L.; Stiftinger, A.; Rechberger, D.; Prangl, B.; Kaufmann, T.; Hiessl, T.; Holly, S.; Heistracher, C.; et al. An Edge Deployment Framework to Scale AI in Industrial Applications. In Proceedings of the 2023 IEEE 7th International Conference on Fog and Edge Computing (ICFEC), Bangalore, India, 1–4 May 2023; pp. 24–32. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.