
Article:
Brintrup, A., Baryannis, G., Tiwari, A. et al. (2025) Trustworthy, responsible and ethical artificial intelligence in manufacturing and supply chains: synthesis and emerging research questions. Data-Centric Engineering, 6. e53. ISSN: 2632-6736

https://doi.org/10.1017/dce.2025.10032

Data-Centric Engineering (2025), 6: e53
doi:10.1017/dce.2025.10032

CAMBRIDGE UNIVERSITY PRESS

SURVEY PAPER

Trustworthy, responsible and ethical artificial intelligence in manufacturing and supply chains: synthesis and emerging research questions

Alexandra Brintrup[1,2], George Baryannis[3], Ashutosh Tiwari[4], Svetan Ratchev[5], Giovanna Martinez-Arellano[5] and Jatinder Singh[6]

[1]Department of Engineering, University of Cambridge, UK
[2]The Alan Turing Institute, British Library, London, UK
[3]Computer Science, University of Huddersfield, UK
[4]The University of Sheffield, UK
[5]University of Nottingham, UK
[6]University of Cambridge, UK
Corresponding author: Alexandra Brintrup; Email: ab702@cam.ac.uk

Received: 15 May 2024; **Revised:** 15 July 2025; **Accepted:** 04 November 2025

Keywords: ethics; Industry 5.0; manufacturing; supply chain; trust

Abstract

In recent years, the manufacturing sector has seen an influx of artificial intelligence applications, seeking to harness its capabilities to improve productivity. However, manufacturing organizations have limited understanding of risks that are posed by the usage of artificial intelligence, especially those related to trust, responsibility, and ethics. While significant effort has been put into developing various general frameworks and definitions to capture these risks, manufacturing and supply chain practitioners face difficulties in implementing these and understanding their impact. These issues can have a significant effect on manufacturing companies, not only at an organization level but also on their employees, clients, and suppliers. This paper aims to increase understanding of trustworthy, responsible, and ethical Artificial Intelligence challenges as they apply to manufacturing and supply chains. We first conduct a systematic mapping study on concepts relevant to trust, responsibility and ethics and their interrelationships. We then use a broadened view of a machine learning lifecycle as a basis to understand how risks and challenges related to these concepts emanate from each phase in the lifecycle. We follow a case study driven approach, providing several illustrative examples that focus on how these challenges manifest themselves in actual manufacturing practice. Finally, we propose a series of research questions as a roadmap for future research in trustworthy, responsible and ethical artificial intelligence applications in manufacturing, to ensure that the envisioned economic and societal benefits are delivered safely and responsibly.

Impact Statement

The presented research is envisioned to lay the groundwork for further research in trustworthy AI for manufacturing, supply chains, and Industry 5.0. Researchers within these areas can consider the provided research questions and roadmap as potential directions to address risks associated with trust, responsibility, and ethics arising from the expanded use of AI. Practitioners in AI applications in manufacturing and supply chains can obtain an understanding of such risks and their impact through illustrative examples that they may have encountered or will encounter in the future. Both researchers and practitioners in data science across different areas of engineering can increase their understanding of trustworthy, responsible, and ethical challenges that may be applicable to their areas of focus.

1. Introduction

Artificial intelligence (AI) has been a major driver of Industry 4.0, with diverse and rich use cases that have helped improve productivity through efficiency gains. Manufacturers increasingly seek AI-based solutions for major challenges the sector is facing; from improving supply chain resilience (Hosseini and Ivanov, 2020), to achieving climate and sustainability goals (Naz et al., 2022). In the UK, 68% of large companies, 34% of medium-sized companies, and 15% of small companies have adopted at least one AI technology, with 44% overall expressing an interest to adopt in the next 3 years (DCMS, 2022).

This rapidly increasing adoption of AI across multiple industrial sectors has also rightfully led to increased scrutiny, moving beyond merely expecting AI to be performant, to requiring that AI solutions are trustworthy, responsible, and ethical. However, the well-documented skills gap in AI within manufacturing training, which is even more pronounced in the case of trustworthy AI development (UK, 2018), has led manufacturing organizations to typically follow one of two scenarios: manufacturing engineers learning how to use AI in an ad hoc manner in response to business requirements, or an AI team with little background in manufacturing tasked with implementation. Both approaches are unable to confidently meet requirements for trustworthiness, leading to poor practices, such as unfair bias in supplier selection, questionable surveillance practices around worker monitoring, failure to retrain models resulting in wrong conclusions, or erroneous maintenance predictions that lead to wasted operational "corrections" (Brintrup et al., 2022).

In a domain as safety critical and vital to the economy as manufacturing, there is a need to ensure the adoption of AI is both safe and appropriate, so that the envisaged societal and economic benefits are delivered responsibly. While there has been extensive policy discussion and survey research on trustworthy AI principles, there is limited understanding of what these mean in the domain of manufacturing and supply chains, and the particular vulnerabilities these domains suffer from during the AI development and deployment cycle. We argue that this thorough understanding can be facilitated through an exploration of current practices in the application of AI technologies that focus on real-world considerations. In particular, we endeavor to respond to the following primary research questions:

1. What are specific trustworthy AI challenges arising in current practices in manufacturing and supply chains?
2. What are the research gaps in trustworthy AI for manufacturing and supply chains that researchers should address?

In this paper, we aim to address these questions through the following research objectives:

1. Conduct a systematic mapping study of policy and research papers that refer to trustworthy AI principles, to identify a frame of reference for our investigation.
2. Collect a set of illustrative cases drawn from current practice that outline the challenges of applying AI responsibly in manufacturing and supply chains.
3. For each illustrative case, identify developments in relevant AI areas that researchers in manufacturing should adopt.
4. Synthesise the insights drawn from each illustrative case and elaborate a research agenda for future studies in the form of research questions that researchers should address.

The contribution of this synthesis and resulting agenda is threefold:

- We offer a real-world view into trustworthy AI challenges arising as a result of current practices in manufacturing and supply chains.
- We identify common risks that arise in the development and deployment of AI solutions in manufacturing and supply chains in relation to trustworthiness.
- We provide directions to focus future research on trustworthy AI in manufacturing and supply chains rooted on issues raised in current practice.

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

In line with the definition proposed by Legg and Hutter (2007) and adapted for supply chain contexts in Baryannis et al., 2018, we define AI as any computational approach that demonstrates the ability to autonomously select and execute actions in pursuit of specific goals while operating in partially unknown or uncertain environments. This broad view encompasses both symbolic (or knowledge-based) AI, such as logic-based reasoning and expert systems, and sub-symbolic approaches, including machine learning, deep learning, and probabilistic modelling. This inclusive interpretation is consistent with the broadest possible AI landscape, from classical symbolic reasoning to generative AI (Sunmola and Baryannis, 2024). Accordingly, the scope of our systematic mapping study and the ensuing research questions apply across the AI spectrum rather than being restricted to data-driven or machine learning-only approaches.

The remainder of this paper is organized as follows. Section 2 describes the methodology followed for the mapping study in relation to trustworthy AI principles and the case-based synthesis and research agenda. Section 3 presents the results of the mapping study, identifying common conceptualizations of trustworthy AI challenges and issues and their relevance to manufacturing. Section 4 then uses a lifecycle-based approach to illustrate how trustworthy AI principles may be impacted when considering the implementation of AI in manufacturing, drawing from 22 illustrative cases. In addition, cross-cutting considerations such as affordability and outsourcing of AI as a service are investigated in Section 5. Finally, Section 6 summarizes the resulting research agenda and outlines relevant future research directions for the manufacturing and supply chains research community.

2. Methodology

In this section, we describe the methodology followed for both parts of this work. Section 2.1 details the methodology followed for the systematic mapping study of trustworthy AI principles, which is based on the work of Petersen et al. (2015). Then, Section 2.2 explains the methodology followed for the case study-based synthesis of trustworthy AI challenges across the AI lifecycle and accompanying research agenda focusing on the manufacturing and supply chains domain.

2.1. Systematic mapping study

The choice of methodology for the review of trustworthy AI principles was determined by the fact that, despite increased recent research around trustworthy, responsible, and ethical AI, this particular research area is still in a nascent stage, compared, for instance, to the more than seven decades of research in the wider field of AI. Hence, a typical systematic literature review would not be appropriate as the aim is not to aggregate evidence on how trustworthy, responsible, and AI practices have been achieved, when such evidence is still not plentiful. In contrast, a systematic mapping study is designed to specifically understand the structure of a research area and the principles that underlie it, rather than gathering and synthesizing evidence (Petersen et al., 2015). As such, it is the appropriate tool to understand and map the range of principles that are driving the development of trustworthy, responsible, and ethical AI solutions.

In the context of software engineering and computer science, an established methodology for conducting systematic mapping studies is that of Petersen et al. (2008) (updated in Petersen et al., 2015). While the methodology follows the general structure of systematic mapping studies as outlined by Petersen et al. (2015), our focus was on synthesizing relevant literature thematically in response to the three guiding research questions, rather than on performing a quantitative mapping of publication trends or keyword frequencies. In the remainder of this section, we outline how we applied this methodology for our exploration of trustworthy AI principles.

The first step in the systematic mapping study process is to define research questions. The following research questions guide our mapping study:

- What are the main concepts of responsibility, ethics, and trustworthiness and their interrelationships?

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

- What are the different concerns arising in relation to algorithmic ethics?
- How do trustworthy AI principles apply to development practices?

The search strategy chosen for the mapping study was an automated search on online databases. As is typical for mapping studies, we used the research questions as a basis for deciding on keywords and opted for a two-level structure, with the two levels conjunctively combined with AND. The first level includes the main focus of the study, AI, and includes the disjunction "artificial intelligence" OR AI. The second level includes the following disjunction of terms related to trustworthy AI: "trustworthy OR responsible OR ethical." Note that we opted for "AI" as a general keyword, rather than specific keywords referring to specific AI algorithms, as this reflects the terminology used by the majority of researchers and practitioners in manufacturing and supply chain disciplines, who frequently adopt the general label 'AI' even when referring to specific subfields.

To maximize coverage, we conducted our search across three major online databases: Scopus, IEEE Xplore, and ACM Digital Library. We chose Scopus over Web of Science based on literature that confirms it has a slightly wider coverage (Pranckuté, 2021). While Scopus does include the majority of IEEE and ACM publications, it is not guaranteed to include all of these, so we chose to search these publishers' databases as well. In terms of time period, we focused on publications from 2016 until today, given that the attention to trustworthy AI is a relatively recent development. The latest searches, the results of which are presented in Section 3 were conducted on 17 February 2025.

The following inclusion criteria were applied: Studies must be written in English and published in journals. Additionally, publications must relate to one or more of the defined research questions by offering a conceptualization of trustworthy AI from a theoretical or practical perspective. In terms of exclusion criteria, we opted not to exclude sources that are not specifically related to manufacturing and supply chains, as this would significantly limit coverage given that trustworthy AI, and AI in general are comparatively less researched in these domains compared to other domains such as healthcare; instead, we provide a brief commentary on the relevance of trustworthy AI to manufacturing in Section 3.5, while the second part of this paper focuses on exploring trustworthy AI challenges that are illustrated through examples within the manufacturing and supply chains domain.

Our search resulted in 3858 articles from Scopus, 238 from IEEE, and 127 from ACM. Following the search, filtering based on inclusion and exclusion criteria was conducted, followed by full-text reading was conducted. Filtering resulted in 48 papers retained. An additional 22 papers were added to the study corpus through snowball sampling, resulting in a corpus of 71 papers. Finally, data extraction and analysis of reviewed publications were conducted independently by two of the co-authors, followed by a consensus meeting, as is common in most mapping studies in the literature (Petersen et al., 2015). Both the search and its results, as well as the mapping study outputs, were reviewed by the last co-author, as a key researcher in trust, compliance, and accountability of emerging and data-driven technologies. Figure 1 presents a flow diagram for the process followed.

Results of the mapping study conducted using the above-described methodology are presented in Section 3, with Sections 3.2, 3.3 and 3.4, respectively, focusing on the first, second, and third research question and Section 3.5 summarizing and relating the study results to manufacturing.

2.2. Synthesis and research agenda

For the second part of our study, we designed a bespoke methodology to structure our analysis and systematically identify the different sources of potential harm to the development of trustworthy AI in the manufacturing and supply chain domain. This methodology is grounded on two main principles: alignment with well-established AI life cycles to ensure complete coverage of all aspects and case study-based analysis to establish links between trustworthy AI development challenges and tangible potential impact on manufacturing and supply chain operations.

The proposed methodology uses an adapted version of the machine learning (ML) life cycle definition suggested by Ashmore et al. (2021). As discussed by Suresh and Guttag (2021), sources of undesirable

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

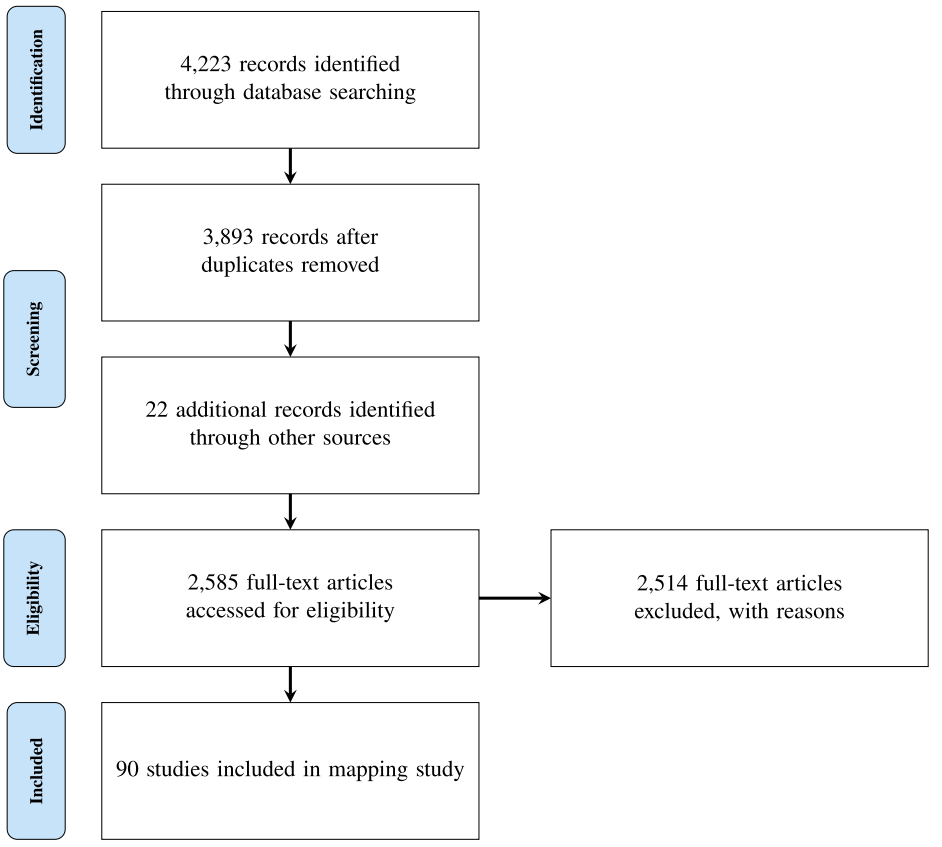


Figure 1. *Flow diagram for the systematic mapping study.*

behavior are not just due to "biased data." The ML pipeline involves a series of choices and practices, including but not limited to model creation, training, and verification, that can lead to undesirable effects. Using Ashmore's life cycle as the basis for our methodology allows us to systematically identify the sources of potential challenges and what they mean within a manufacturing and supply chain context. For each phase in the life cycle and each activity within each phase, we explore a wide variety of issues that relate to trustworthy AI. In addition to the synthesis of trustworthy AI challenges, which is presented in Section 4, we go beyond individual phases and activities and explore cross-cutting considerations as well, in Section 5.

As far as ensuring that our exploration of trustworthy AI issues is linked to their tangible impact on manufacturing and supply chain operations, we opted to make use of illustrative examples that explain how each issue raised may result in a given trustworthiness concern. Due to the limited research and development of trustworthy AI solutions in the manufacturing and supply chain domains, the provided examples are drawn from expert knowledge and industry experience.

Illustrative examples are also then coupled with one or more research questions that identify potential areas for research to address the challenges identified and mitigate their impact. Research questions were derived through abductive reasoning informed by illustrative examples of real-world industrial practice encountered through the authors' ongoing engagements with AI deployment in manufacturing and supply chain contexts. These research questions arise from challenges encountered in articulating and applying ethical and trustworthy AI principles across different industrial stages, from initial scoping and design, to procurement, implementation, deployment, and assurance. They also reflect recurring concerns around bias, fairness, explainability, accountability, risk management, and regulatory uncertainty. The questions

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

were designed to reflect the recurring themes observed in these engagements, while remaining broad enough to allow mapping to the wider academic and policy literature. This structure enables the study to bridge theoretical formulations with practical concerns emerging in applied settings. Collectively, these research questions constitute our proposed research agenda for future exploration of trustworthy AI in the context of manufacturing and supply chains. A summary of the results of the second part of our study is presented in Table 2.

3. Trustworthy, responsible and ethical AI: a systematic mapping study

In this section, we present the results of the systematic mapping study on trustworthy AI principles conducted following the methodology in Section 2.1. Our exploration of sources published over most of the past decade confirms that the accelerated growth of the AI field and the increased adoption of algorithmic systems have led to a growing concern regarding the impact, implications, and consequences of AI-driven systems. Some of these issues include: amplification of bias, loss of human privacy, use of AI to create digital addiction, social harms caused by digital surveillance and criminal risk assessment, disinformation through fake text generated by AI, and loss of employment or quality of employment as machines replace humans (Brundage et al., 2020). Researchers and policy makers have warned that significant efforts should be devoted to ensuring the use of AI is in the public interest, that it works for society, and is not detrimental to humanity and human well-being. As we write this paper, there are calls from AI industry leaders themselves to embargo major AI releases by six months to evaluate unintended consequences (Hern, 2023).

Much work is underway toward such concerns in a broader context, from policy makers designing regulatory frameworks to academic research proposing foundational principles for ethical, responsible, and trustworthy AI. This has yielded a multitude of frameworks that encourage structured, systematic exploration. Notable examples include: the assessment list for trustworthy AI set up by the European Commission's High-Level Expert Group on AI (High-Level Expert Group on Artificial Intelligence, 2019); the EU AI Act (European Commission, 2021); the AI Bill of Rights Blueprint by the United States White House Office of Science and Technology Policy (Office of Science and Technology Policy, 2022); and the AI Risk Management Framework by the National Institute of Science and Technology (NIST) (Trustworthy and Responsible AI Resource Center, 2023). In addition, efforts have been made to consolidate information on these issues from across different sources. For example, Algorithm Watch provides an assessment of more than 170 automated decision making systems in Europe in their 2020 Automating Society Report (Chiusi, 2020); Jobin et al. (2019) identifies 84 documents outlining different principles; and Newman (2023) introduces a taxonomy of 8 characteristics and 150 properties of trustworthiness for AI, drawn from a comprehensive analysis of the landscape of trustworthy AI. Note, however, that principles articulated by Western academics and technology providers are not necessarily representative globally (Brundage et al., 2020). For example, a deeper investigation showed that Beijing AI Principles show disagreements between Western and non-Western AI principles, despite them using the same terminology (Paleyes et al., 2022). It is also worth noting that practitioners struggle with implementing these high-level frameworks and regulatory guidance is missing.

While studies on the rise of AI in the manufacturing sector have been widely noted, there has been limited work on the risks that it may raise in a manufacturing organization or a set of organizations in a supply chain. One notable exception is the work arising from the Horizon 2020-funded ASSISTANT project, yielding the Trustworthy AI Project Risk Management Framework (TAI-PRM) by Besinger et al. (2024), and the Responsible AI (RAI) framework by Vyhmeister and Castane (2025), which are both explored in our mapping study. Further work needs to be undertaken in both interpreting principles related to trustworthy, responsible, and ethical AI within a manufacturing and supply chain context, and also for ensuring that they are upheld ubiquitously.

It should be noted that the conducted systematic mapping study differs from a standard systematic literature review in that it does not attempt to provide a thorough review of all existing proposals related to trustworthy AI. Rather, we aim to map the broader discourse around trustworthy, responsible, ethical AI as

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

well as explainable AI to provide an indication of the common themes considered in this space, so as to provide a basis for linking these to the challenges of AI in the context of manufacturing and supply chains. Further, for the purposes of this paper, we use the term "trustworthy AI" in the remainder of this paper to refer collectively to the themes covered across the spectrum of these terms.

3.1. Bibliometric analysis

To provide a broader overview of the trustworthy AI literature, we supplemented the systematic mapping study with a bibliometric analysis using VOSviewer (van Eck and Waltman, 2010). This analysis aimed to explore recurring themes, geographical research activity, and patterns of scholarly collaboration across the included corpus. Four visualizations were generated: one based on keyword co-occurrence (Figure 2, one on country-level co-authorship (Figure 3), and two on author collaboration networks (Figure 4).

The keyword co-occurrence map (Figure 2) offers an overview of the main topics addressed in the literature. The central position of terms such as artificial intelligence, machine learning, and ethics reflects their foundational role in this domain. Around these core concepts, several distinct clusters are visible. For instance, terms like explainability, accountability, and fairness form a tightly connected group that corresponds to ongoing discussions around ethical principles in AI. Other terms, such as governance, human–computer interaction, and philosophical aspects, indicate the field's interdisciplinarity and engagement with policy and sociotechnical issues. The visualization highlights the prominence of technical concepts such as explainable AI alongside more normative concerns, underscoring the hybrid nature of current research in this area.

Figure 3 illustrates the geographical distribution of authorship and co-authorship. The most active countries in terms of publication volume and collaboration include the United Kingdom, the United States, Germany, and Italy. A relatively dense cluster of European countries is visible, suggesting strong regional collaboration, while countries such as China, Australia, and Canada appear as active contributors with varying degrees of international co-authorship. Several countries from the Global South, including Nigeria and Egypt, are represented but relatively disconnected. This visualization draws attention to the

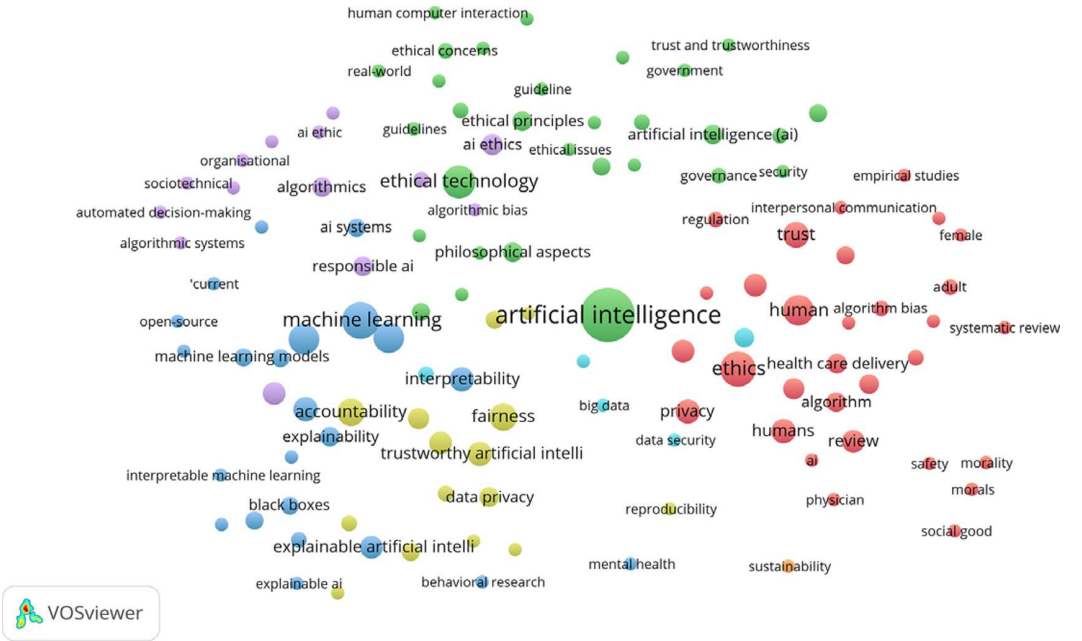


Figure 2. *Keyword co-occurrence network visualizing the conceptual structure of trustworthy/responsible AI literature. Node size reflects term frequency; colours indicate thematic clusters.*

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

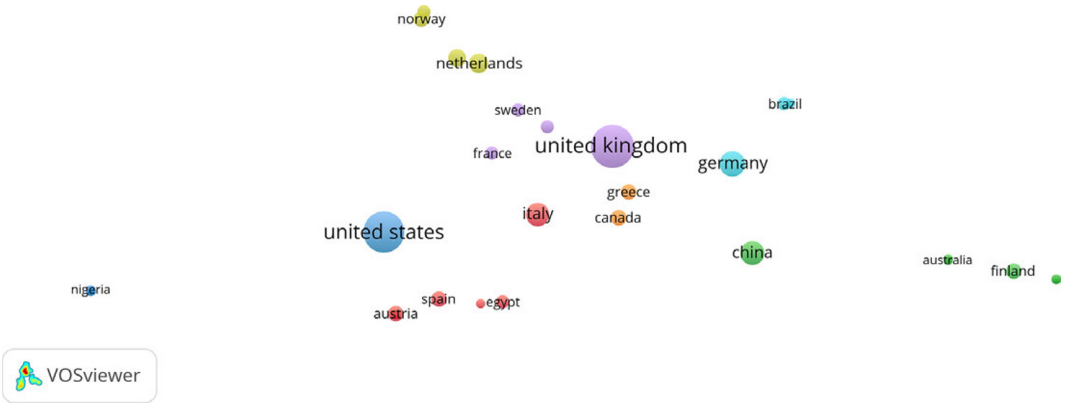


Figure 3. *Geographical distribution of research contributions to the trustworthy AI literature. Node size indicates publication volume; proximity indicates collaboration strength.*

asymmetries in global participation and highlights the importance of ensuring broader inclusion in international AI governance debates.

Figure 4a and b present author-level co-authorship networks. The first (Figure 4a) shows the largest connected component, centered on a group of researchers who frequently collaborate on issues such as algorithmic bias, transparency, and AI governance. The second (Figure 4b) extends this view to include all clusters with link strengths greater than 1. This wider perspective reveals the fragmentation of the field into several relatively distinct communities, each with its own internal collaborations but limited inter-group connectivity. The lack of cross-cluster ties suggests that, although the field is active and collaborative, there may be missed opportunities for interdisciplinary integration—particularly between technical, legal, and social science perspectives on trustworthy AI.

3.2. Relationships among responsibility, ethics, and trustworthiness

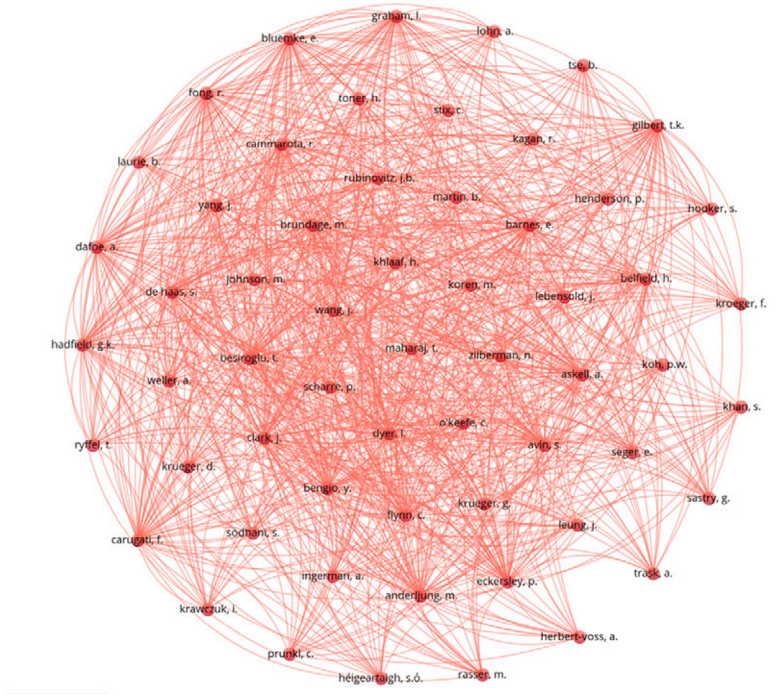
With regard to the first research question of the mapping study on concepts of responsibility, ethics, and trustworthiness, one common approach in the literature, which we follow in our analysis, is to view responsible and ethical requirements as being fundamental prerequisites to trusting an AI system. According to Smuha (2019), the European Union defines trustworthy AI as being "lawful (respecting all applicable laws and regulations), ethical (respecting ethical principles and values) and robust (both from a technical perspective while taking into account its social environment)." Responsibility in the form of accountability is defined as one of seven key requirements that AI systems should meet in order to be deemed trustworthy, encompassing responsible development, deployment, and use.

In his analysis of reliable, safe, and trustworthy AI, Shneiderman (2020) primarily views responsibility from the perspective of clarifying the role of humans in AI failures, also mentioning responsibility in combination with fairness and explainability as goals of human-centred AI. Kaur et al. (2022) defines accountability/responsibility as one of the requirements for trustworthy AI, alongside fairness, explainability, privacy, and acceptance.

Thiebes et al. (2021) posits that "AI is perceived as trustworthy by its users (e.g., consumers, organizations, society) when it is developed, deployed, and used in ways that not only ensure its compliance with all relevant laws and its robustness but especially its adherence to general ethical principles." For the latter, they adopt the following ethical principles: beneficence, non-maleficence, autonomy, justice, and explicability. Responsibility is only considered as an aspect of explicability and justice, in the sense of holding someone legally responsible in case of an AI failure.

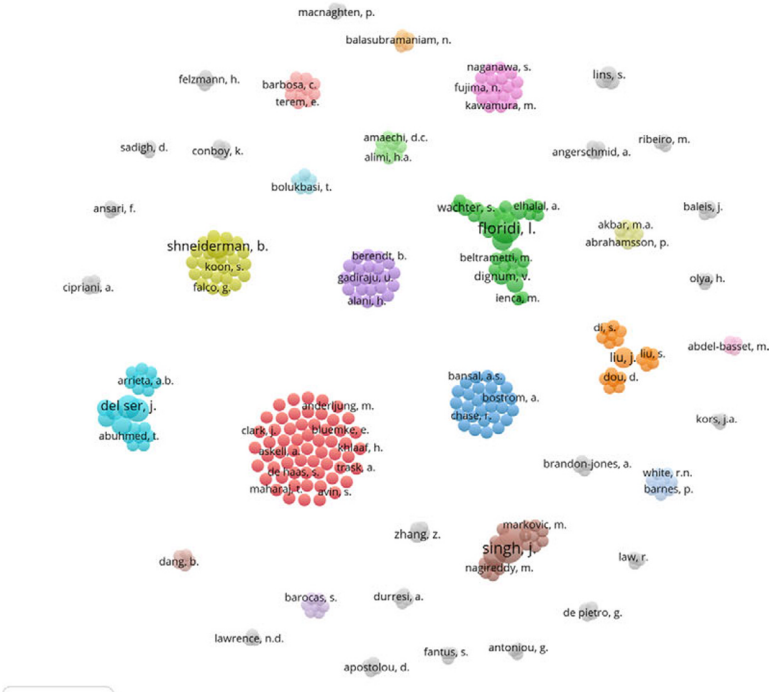
Trustworthiness is also at the core of the recently published white paper of the UK Government on AI regulation (Department for Science, Innovation and Technology, 2023). One of the three main aims of the

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press



VOSviewer

(a) Largest author cluster



VOSviewer

(b) All clusters where link strength > 1

Figure 4. *Author co-authorship networks in the trustworthy AI literature.*

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

proposed regulatory framework is to increase public trust in the use and application of AI and is underpinned by five principles: safety, security and robustness, appropriate transparency and explainability, fairness, accountability and governance, and contestability and redress.

Finally, Newman (2023) also places trustworthiness at the centre of discussions around responsibility and ethics, and develops a comprehensive taxonomy of 150 trustworthiness properties. These properties relate to one of the following eight trustworthiness characteristics based on NIST's AI Risk Management Framework (Trustworthy and Responsible AI Resource Center, 2023): valid and reliable, safe, secure, and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, fair with harmful bias managed, and responsible practice and use. This taxonomy clearly positions responsibility and ethics as contributors and prerequisites to trustworthiness, rather than outcomes of it.

Our mapping confirms that this view remains prevalent in recent literature. For example, Mentzas et al. (2024) and Kattnig et al. (2024) reaffirm the framing of responsibility and accountability as components of trustworthiness, while authors such as Stahl (2023) and Law et al. (2025) explore how the broader ecosystems and societal domains in which AI operates shape the interpretation of responsibility and trust. A number of sources continue to treat ethical principles such as justice and autonomy as foundational to earning trust in AI, especially in domain-specific applications such as healthcare (Zhang and Zhang, 2023; Ueda et al., 2024).

Note, however, that an alternative viewpoint is to cast trustworthiness as a prerequisite for responsible AI. Wang et al. (2020) groups responsible for AI practices into four categories: training/education, risk control, ethical design, and data governance. Trust building is one component of data governance, along with explainability and transparency. The narrative provided is a quite narrow view of trust that is centred around reducing bias through high-quality data and ensuring there is consent for sharing data.

Arrieta et al. (2020) defines seven responsible AI principles: explainability, fairness, privacy, accountability, ethics, transparency, security/safety. Trust is not included independently, but rather shown as an aspect or goal of explainability. As the authors explain, trustworthiness and explainability are not equivalent, as being able to explain outcomes does not imply that they are trustworthy, and vice versa.

Rather than viewing one as a prerequisite of the other, some researchers place both responsibility and trustworthiness at the same level, as principles of ethical AI. A comprehensive review of AI guidelines in the literature is conducted by Jobin et al. (2019), producing the following list of ethical AI principles in order of commonality: transparency, justice/fairness, non-maleficence, responsibility /accountability, privacy, beneficence, freedom /autonomy, trust. Responsibility and accountability are rarely defined, but recommendations focus on "acting with integrity and clarifying the attribution of responsibility and legal liability." Trust is referenced in relation to customers trusting developers and organizations and trustworthy design principles.

3.3. Algorithmic ethics

We now focus our attention on the second research question around "ethical algorithms." Mittelstadt et al. (2016) developed a map with different types of ethical concerns useful for doing a rigorous diagnosis of ethical concerns emerging from AI, which are used to evaluate ethical outcomes in AI applications. In their map, "inconclusive evidence" refers to the data analysis stage where results produce probabilities but also uncertain knowledge. Here authors point out cases where correlations are identified but the existence of a causal connection cannot be posited. Failure to recognize this might then lead to unjustified actions. "Inscrutable evidence" refers to a lack of transparency regarding both the data used to train an algorithm and a lack of interpretability of how data-points were used by an algorithm to contribute to the conclusion it generates. This is commonly referred to as the "black-box" issue, leading to non-obvious connections between the data used and the resulting conclusions. 'Misguided evidence' refers to the fact that an algorithmic output can never exceed the input and thus conclusions can only be as reliable and neutral as the data they are based on, which can lead to biases. 'Unfair outcomes' refer to actions that are based on conclusive, scrutable, and well-founded evidence, but they have a disproportionate, disadvantageous impact on one group of people, often leading to discrimination. "Transformative effects" refer to

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

algorithmic activities, such as profiling the world by understanding and conceptualizing it in new, unexpected ways, triggering and motivating actions based on the insights it generates (Morley et al., 2020). This can lead to challenges for autonomy and informational privacy.

Ethics by design include best practices in the development of AI to mitigate the above group of concerns—for example, the establishment of an ethics board (Leidner and Plachouras, 2017) and integration of "ethical decision routines in AI systems" (Hagendorff, 2020), whereby decision algorithms are explicitly designed to respect ethical values.

Many papers in our extended mapping continue to reflect and expand on these concerns. Hagendorff (2020) provides a critical assessment of over 20 ethical AI guidelines, identifying key omissions and inconsistencies in how algorithmic harms are conceptualized. Kazim and Koshiyama (2021) and Lewis and Marsh (2022) explore how algorithmic manipulation and opacity undermine human agency and trust, while Wang et al. (2023) and Zhang and Zhang (2023) document the emergence of ethical risks in clinical AI systems, such as automation bias and unequal access to treatment.

A number of studies stress the importance of domain context when considering algorithmic ethics. For instance, Giovanola and Tiribelli (2023) argue that standard definitions of fairness are insufficient for healthcare settings and propose a "relational fairness" approach that considers patient dignity and systemic inequalities. Similarly, Nguyen et al. (2023) and Tang et al. (2023) examine stakeholder perceptions of fairness and transparency in education and medical AI, respectively, identifying mismatches between technical fairness mechanisms and societal expectations.

Several contributions also examine the socio-technical nature of algorithmic ethics, calling for a shift from principle-level discussions to implementation frameworks. Starke et al. (2022) conduct a systematic review of perceived algorithmic fairness and note that technical fixes alone are unlikely to address public concerns. Others, such as Radanliev (2024) explores the use of privacy-preserving technologies like federated learning and homomorphic encryption as a means of embedding ethics into algorithmic architecture.

3.4. Developing trustworthy AI

In contrast to the first two areas of the mapping studies, which focused on principles underlying trustworthy AI research, the third area looks at trustworthy AI from the point of view of development. Building on principles, practitioners must consider, undertake and employ various measures and safeguards so as to mitigate the risks of the technology, such that they consider and address the various concerns to avoid negative consequences on human and societal well-being (Dignum, 2023).

Toward this, there have been various approaches describing responsible development practices. For example, Arrieta et al. (2020) and Sambasivan and Holbrook (2018) describe responsible AI as being concerned with the design, implementation and use of ethical, transparent, and accountable AI technology in order to reduce biases, promote fairness, equality, and help facilitate interpretability and explainability of outcomes.

When conceptualizing responsible AI, the principles of responsible research and innovation (RRI) have served as a starting point to consider and anticipate the consequences of a particular technology in society (Owen et al., 2012). Shaped by contributions from Science and Technology Studies, this approach has been established and prominent in recent projects funded by the European Commission. The application roadmap of responsible AI includes continuous reflection on context and civil society, such as third sector organizations, so as to align the AI development process and outcomes with society's expectations. Decision processes must be visible and transparent to ensure that developers are on track regarding their responsibilities, and the development process must allow users and stakeholders of technologies to criticize outcomes.

In terms of areas of consideration specific to the AI lifecycle, the topics of transparency (including explainability and interpretability) and fairness (bias) have received considerable attention by the technical and engineering communities, which we explore below. Note, however, that there is a clear realization that issues of trustworthy AI are inherently socio-technical (Kroll et al., 2017; Raji et al., 2020),

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

and require a consideration of technical, organizational, and human processes aspects, throughout the technology development, operation, and use (Cobbe et al., 2021) as well as their supply chains (Cobbe et al., 2023).

Recent literature extends this understanding by introducing development frameworks tailored to specific trustworthiness requirements. For example, Thiebes et al. (2021) propose the DaRe4TAI framework, which operationalizes five trustworthiness principles (including autonomy and justice) through a series of design dimensions. Similarly, Mentzas et al. (2024) provide a structured overview of methods and toolkits for implementing trustworthy AI in practice, including risk assessment, fairness evaluation, and transparency monitoring.

The use of standardized development guidance is also highlighted in technical and regulatory frameworks. Several papers refer to the NIST AI Risk Management Framework and the assessment list for trustworthy artificial intelligence (ALTAI) as concrete tools to translate ethical principles into engineering practices (Díaz-Rodríguez et al., 2023; Radclyffe et al., 2023; Kattnig et al., 2024). These studies suggest that development practices must be embedded across the AI lifecycle, from data preparation to post-deployment auditing.

Some contributions also emphasize the limitations of high-level frameworks in practice. Morley et al. (2023) find that developers and policymakers often struggle to operationalize principles such as fairness and accountability without domain-specific guidance. Stahl (2023) calls for a shift toward ecosystem-level responsibility, recognizing that AI trustworthiness depends on the combined actions of system developers, deployers, regulators, and affected communities.

Finally, we find an increasing number of domain-specific development strategies that adapt general trustworthiness principles to localized contexts. For instance, Ueda et al. (2024) introduce the FAIR framework for healthcare AI development, and Law et al. (2025) propose a staged AI evolution trajectory model tailored for the hospitality sector. These approaches illustrate the growing maturity of trustworthy AI discourse, which is moving from abstract discussion to actionable development practices.

3.4.1. Transparency

Explaining AI decisions and interpreting model outputs is commonly included in the discussion of responsible, ethical, and trustworthy AI. Explainability and interpretability are often used interchangeably in the literature (Arrieta et al., 2020). As argued by Antoniou et al. (2022), interpretability has a narrower focus that primarily relates to the degree to which ML model outputs can be interpreted in relation to relevant data. Explainability builds on model interpretability by including explanations that are not exclusively related to data and ML but may relate to expert knowledge and other psychological, cognitive, or philosophical aspects (Adadi and Berrada, 2018). An explainable approach is one that allows for identifying a complete reasoning pathway from input to output.

Three main aspects of ML interpretability are recognized in literature (De Laat, 2018). Ex-ante refers to how an algorithm arrived at a decision, offered in the form of a description of the inner working of the models, including what is the working procedure of an algorithm and how it generally processes input data to produce output. Ex-post refers to which training data has been used to derive results, highlighting which set of evidence/training data has been used to make each decision. A third aspect focuses on metrics used to measure the validity of the result. Here, uncertainty measures are often used, allowing users to determine confidence intervals and help them decide whether the model has made a valid decision.

It is also important to consider the role transparency plays in its broader context, raising questions about what sort of transparency, and to whom and for what. In practice, transparency generally will not solve issues with the technology (Ananny and Crawford, 2018), but can provide a basis for supporting recourse, repair, and accountability more generally (Cobbe et al., 2021; Williams et al., 2022). Transparency is also a key risk consideration in the framework developed by Vyhmeister and Castane (2025) and refers to the provision of adequate documentation and provenance processes, as well as periodic performance reports.

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

3.4.2. Fairness

Fairness refers to biases in data and deployment, which can lead to systematic disadvantages for marginalized individuals and groups. This requirement advises that AI development cycles should include methods for checking AI bias in data and decision-making processes. Fairness is included as one of the six components of responsible AI by Besinger et al. (2024), focusing on addressing issues of discrimination and bias and promoting inclusivity, and including the aspiration of developing algorithms that are fair by design.

Many open-source ML "fairness toolkits" have been developed to assist ML practitioners in assessing and addressing unfairness in the ML systems they develop (Wexler et al., 2019). For instance, companies such as Microsoft, Google, and IBM have published combinations of toolkits and guidelines that incorporate fairness. Fairness also features prominently as one of the 6 components of the RAI framework by Besinger et al. (2024).

Recent studies have shown that practitioners need more practical guidelines from fairness toolkits in order to be able to contextualize ML fairness issues and communicate them to non-technical colleagues (Lee and Singh, 2021; Deng et al., 2022). Deng et al. (2022) identified four design requirements ML practitioners had when using fairness toolkits: the ability to use the toolkit to learn more about ML fairness research, rapid use due to time constraints, the ability to integrate toolkits into existing ML pipelines, and using toolkit code repositories to implement ML fairness algorithms.

3.4.3. Requirements beyond the Newman taxonomy

While the eight characteristics of trustworthiness proposed by Newman (2023) provide a comprehensive and structured lens, our expanded mapping also identified additional requirements and conceptual refinements that do not fit neatly into this taxonomy but are prominent in recent literature.

One key extension is the notion of *relational fairness*, as introduced by Giovanola and Tiribelli (2023). This concept critiques conventional fairness definitions focused on statistical parity or equal treatment and instead highlights the importance of context, dignity, and care, especially in healthcare settings. Similarly, Felzmann et al. (2019) argue for *relational transparency*, noting that the informational needs of users vary by context and stakeholder role, and that one-size-fits-all explanations may hinder, rather than help, trust and accountability.

Another cluster of extensions relates to broader socio-political dimensions of AI governance. Paraman and Anamalah (2023) call for the inclusion of *sovereignty*, *human-in-command*, and *autonomous sustainability* as guiding principles, arguing that responsible AI must consider collective values and long-term planetary impacts. Meanwhile, Stahl (2023) advocates for *ecosystem-level responsibility*, which acknowledges the multi-actor nature of AI systems and emphasizes that responsibility must be shared across designers, deployers, regulators, and affected communities. In the same context, Besinger et al. (2024) argue in favor of human-centric AI as one dimension of responsible AI, highlighting the need for centering on human values and supporting human agency, ensuring alignment with human dignity and interests.

A related domain-specific perspective is offered by Law et al. (2025), who highlight the importance of preserving human values in the evolution of AI in hospitality and tourism. They warn that technological solutions can disrupt emotionally sensitive experiences and propose *AI-human experience preservation* as a complementary dimension to existing trustworthiness criteria.

Finally, Besinger et al. (2024) and Vyhmeister and Castane (2025) bring in green and environment-related requirements into scope. Besinger et al. (2024) refers to green AI as both the requirement to making AI execution and deployment sustainable and energy-efficient and the requirement to harness AI to address environmental problems, such as climate change and resource depletion. Vyhmeister and Castane (2025) refers to the former as "environmental well-being."

Together, these perspectives suggest that while Newman's taxonomy remains a valuable reference point, additional dimensions may need to be considered to fully capture the values and risks that arise in diverse AI deployment contexts.

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

3.5. Trustworthy AI principles and their relevance to manufacturing

Table 1 summarizes the results of the mapping study in the form of the most common requirements for trustworthy AI proposed in the literature. Although some authors use some of these terms to encompass other requirements, the taxonomy by Newman (2023) provides a framework that encompasses most aspects in literature.

As we see from the above terminology, there is considerable work pertaining to responsible, trustworthy, and ethical AI principles and these often overlap, with debate taking place over their specific taxonomy. Defining such a taxonomy is an important area of deliberation, as it allows for structured thinking. However, for the purposes of this paper, we refrain from strictly defining this fluid field and opt to be as broad and inclusive as possible. For this reason, we use the term "trustworthy AI" as an umbrella term to encompass principles of ethical, responsible, and trustworthy AI.

We argue that manufacturers need to invest in all trustworthy AI principles, and not just a subset of them, when considering how they develop, deploy, and practice AI in their organizations. Figure 5 illustrates how each principle in the taxonomy by Newman (2023) has an impact on different aspects of manufacturing.

Manufacturing companies do not exist in isolation. They impact not only the profitability of stakeholders but also the well-being of their workers. The products and services that manufacturing engineers design and produce impact society. The use of natural resources and waste that is generated during production and delivery has a profound impact on the environment, and decisions companies make on suppliers can have wide wide-reaching impact on global economies. Therefore, it is crucial that manufacturing adopts AI in a lawful and ethical manner that is robust, safe, and avoids negative consequences to human society and well-being. It is important to be able to prove that an organization does so, via algorithms and datasets that can be scrutinized. Given its wide-ranging remit, we thus feel that the field of manufacturing needs to be inclusive when thinking about trustworthy AI principles. Thus, in the remainder of this paper, we shall review specific challenges that manufacturing must face in order to be able to create and deploy trustworthy AI in its broadest sense.

4. Trustworthy AI challenges across the AI lifecycle

The previous section discussed different, related terminology when considering trustworthiness of AI. While the frameworks associated with these concepts provide a useful starting point in understanding how AI can affect industrial contexts adversely, practitioners often highlight that they remain too general and abstract for any useful insight to be gained from them (Shneiderman, 2021; Trocin et al., 2021).

As introduced in Section 2.2, we thus propose a process-oriented lens for analyzing challenges in relation to AI trustworthiness in the manufacturing context, in order to connect concerns raised in the trustworthy AI community to the development and implementation processes that may result in their emergence. While Ashmore's ML life cycle definition mainly encompasses statistical AI paradigms, in further subsections, we will discuss the implications of trustworthy AI in both symbolic and sub-symbolic AI approaches. For example, phases such as knowledge elicitation and rule encoding in expert systems align with the data collection and training stages, respectively. To facilitate this presentation, "data" in the remainder of this section refers to both data as used in a typical statistical AI approach, as well as data in the form of expert knowledge. Moreover, "model" refers to any intelligent model, ranging from ML models to knowledge models. We also provide additional commentary specific to knowledge-based AI where necessary, to clarify any differences in relation to data-driven AI.

4.1. Data management

Data management focuses on the preparation of datasets needed to build an AI model, which typically include data collection, augmentation, and pre-processing processes.

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

Table 1. *Mapping of literature to Newman's AI trustworthiness requirements*

Requirements	Literature
Valid and reliable	Mittelstadt et al. (2016) [conceptualization], Shneiderman (2020), Kaur et al. (2022), Newman (2023), Floridi et al. (2020), Ntoutsi et al. (2020), Bostrom et al. (2024)
Safe	Smuha (2019) [robustness, reliable, accurate, reproducible and safe], Floridi (2019) [robustness and safety], Arrieta et al. (2020) [security], Newman (2023), Shneiderman (2020), Shneiderman (2021), Díaz-Rodríguez et al. (2023), Elendu et al. (2023), Mennella et al. (2024), C. Wang et al. (2023), Zhang and Zhang (2023), Vyhmeister and Castane (2025) [robustness]
Fair with Harmful Bias Managed	Mittelstadt et al. (2016), Jobin et al. (2019) [justice], Smuha (2019), Floridi (2019), Arrieta et al. (2020), Thiebes et al. (2021) [justice], Shneiderman (2020), Kaur et al. (2022), Newman (2023), Martin (2019), Ntoutsi et al. (2020), Hermann (2022), Kordzadeh and Ghasemaghaei (2022), Starke et al. (2022), Ali et al. (2023), Díaz-Rodríguez et al. (2023), Giovanola and Tiribelli (2023) [relational fairness], Nguyen et al. (2023), Tang et al. (2023), C. Wang et al. (2023), Zhang and Zhang (2023), Ueda et al. (2024), Kattnig et al. (2024), Law et al. (2025), Besinger et al. (2024)
Secure and resilient	Mittelstadt et al. (2016) [manage uncertainty], Newman (2023), Radanliev et al. (2024), Besinger et al. (2024)
Explainable and interpretable	Arrieta et al. (2020), Shneiderman (2020), Y. Wang et al. (2020), Kaur et al. (2022), Newman (2023), Markus et al. (2021), Ding et al. (2022), X. Li et al. (2022), Ali et al. (2023), Balasubramaniam et al. (2023), Joyce et al. (2023), Marcinkevičs and Vogt (2023), Ueda et al. (2024), Besinger et al. (2024), Vyhmeister and Castane (2025)
Privacy-enhanced	Jobin et al. (2019), Smuha (2019), Floridi (2019) [privacy and data governance], Arrieta et al. (2020), Kaur et al. (2022), Newman (2023), Mennella et al. (2024), Radanliev et al. (2024), Besinger et al. (2024), Vyhmeister and Castane (2025)
Accountable and transparent	Jobin et al. (2019), Smuha (2019) [explainability and traceability], Floridi (2019), Arrieta et al. (2020), Y. Wang et al. (2020) [Data Governance], Kaur et al. (2022), Newman (2023), Mittelstadt et al. (2016) [risk assessment, traceability], Floridi and Taddeo (2016), Martin (2019), Felzmann et al. (2019) [relational transparency], Hagendorff (2020), Ryan and Stahl (2021), Corrêa et al. (2023), Khan et al. (2023), Nguyen et al. (2023), Tang et al. (2023), Zhang and Zhang (2023), Bostrom et al. (2024), Kattnig et al. (2024), Laux et al. (2024), Mentzas et al. (2024), Ueda et al. (2024), Radclyffe et al. (2023), Besinger et al. (2024), Vyhmeister and Castane (2025)
Responsible practice and use	Jobin et al. (2019) [inc. beneficence, non-maleficence, freedom and autonomy], Smuha (2019) [lawful], Floridi (2019) [human agency and oversight, societal and environmental well-being], Arrieta et al. (2020) [ethics], Thiebes et al. (2021) [inc. beneficence, non-maleficence, freedom and autonomy], Shneiderman (2020) [ethical], Newman (2023), Floridi and Taddeo (2016), Floridi et al. (2018), Hagendorff (2020), Ryan and Stahl (2021), Thiebes et al. (2021), Ozmen Garibay et al. (2023), Díaz-Rodríguez et al. (2023), Khan et al. (2023), Morley et al. (2023), Stahl (2023), B. Li et al. (2023), Mentzas et al. (2024), Law et al. (2025)

Note. New additions are shown in blue.

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

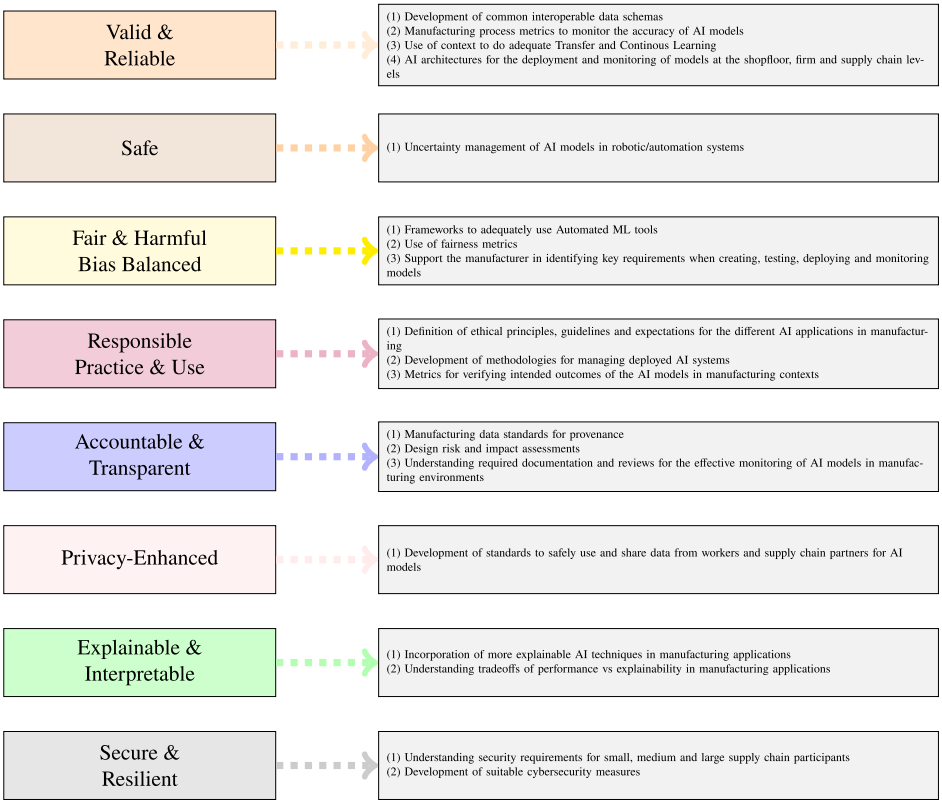


Figure 5. *Mapping key principles in Newman (2023) to corresponding needs in the manufacturing context.*

4.1.1. Data collection

Data collection involves activities that aim to discover and understand what information is available, as well as ensuring that this information is easily accessible and processable. The task of discovering which data exists and where in an organization is usually a challenge by itself, especially in large manufacturers with multiple facilities and geographic locations. Companies often do not know the full extent of data and knowledge they have, and what it can be used for. Data may be dispersed in emails of individuals, physical or digital documents, in legacy systems, Supervisory Control and Data Acquisition (SCADA), Manufacturing Execution Systems (MES), and Product Lifecycle Management (PLM) systems, as well as structured databases such as Enterprise Resource Planning (ERP) or Customer Relationship Management CRM) bases. In addition to these internal data, manufacturers are increasingly looking into leveraging publicly available datasets, ranging from weather and traffic forecasts to social media.

Interoperability. When identifying different data sources relevant to the AI problem at hand, a key challenge is that these sources may have different schemas, formatting conventions, and differing storage and access requirements. For example, downloading social media information may involve a different procedure than simply downloading an ERP snapshot. Joining this information into a single dataset suitable for analysis is sometimes referred to as the data integration process. Organizations may want to trial pilots using representative datasets before setting up projects that allow easily repeatable access and integration mechanisms, which itself is determined by the usefulness of the pilot and its cost.

Addressing data integration and interoperability issues has long been the focus of research efforts in relation to information systems, especially since the proliferation of big data Kadadi et al. (2014). Standardization through commonly agreed terminologies and taxonomies is most often suggested as a

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

Illustrative Example 1: Two food manufacturers are collecting data for their products through Food Product Information Forms (FPIF), which differ in both content and structure. The manufacturers enter into an agreement that involves combining products and ingredients, and need to be able to exchange data for their products and store it in a jointly managed repository. Given the differences in each manufacturer's FPIF, issues are raised with regard to data entries with different names that refer to the same concept, and data entries that are unique to one of the FPIFs. Failing to address these issues may compromise the validity of models produced based on these data.
Risk: Missing data and wrong model, potentially leading to inscrutable evidence.

solution to these issues, leveraging technologies relying on data schemas, such as those based on XML, and semantic web technologies and ontology languages, such as RDF and OWL (Pauwels et al., 2017). In this context, the following question arises:

- **RQ 1:** How can data producers and owners be supported in implementing common data schemas and knowledge models to improve interoperability?

Provenance. ML relies on learning models based on datasets, while knowledge-based AI relies on models built based on expert knowledge. In both cases, there is a common key challenge of identifying the sources of data and/or knowledge and of determining whether these sources can be trusted.

Illustrative Example 2: A food manufacturer is developing an automated system for determining food allergens in its food products. The system relies on a knowledge graph that is built based on food allergen information provided by other food manufacturers for each ingredient. An unreported or misreported food allergen on their part can lead to an incomplete or incorrect knowledge graph. This, in turn, may lead to unreported allergens on food product labels, which can lead to significant consequences.
Risk: Missing or incorrect data leading to unsafe model operation, inscrutable evidence.

The challenge of provenance has been well-researched in the field of knowledge engineering. Provenance of a resource is defined by the W3C Provenance Incubator Group (2010) as "a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource." In this context, the PROV family of documents has been developed, which includes PROV-O, an ontology allowing to attach provenance information on a knowledge model. PROV-O includes three main aspects: entities, activities, and agents, capturing information on agents performing activities on entities. In the standard provenance use case, this allows recording the person or company that is the source of a particular piece of knowledge.

PROV-O and related research provide the infrastructure and mechanisms to capture provenance information and set the foundations of trusting information by knowing and trusting its source. However, to ensure trustworthiness, the knowledge acquisition and engineering processes need to include time and effort spent on using provenance infrastructure and mechanisms to record the necessary information. This has been described in various contexts, including for AI systems (Pasquier et al., 2018; Huynh et al., 2021), both within and across their organizations and supply chains (Singh et al., 2019), yet tends to be mostly conceptual with many opportunities for future work. This is captured in the following question:

- **RQ 2:** How can provenance mechanisms be leveraged in association with data and knowledge acquisition processes in a manufacturing context?

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

Bias. A further challenge during data collection is the inadvertent introduction of bias in the collected data (Suresh and Guttag, 2021). Most AI models rely on historical data to make decisions, which means if they are developed based on data that contains hidden biases, their decisions will be biased, despite the irony that many times, ML, in particular, is marketed as an approach to remove human error and bias from a manufacturing use case.

Data collection issues resulting in biased data would result in inconclusive evidence by suggesting spurious correlations, unfair outcomes, as data errors may result in disproportionate impact on one group of people or organizations, and this effect may even be difficult to detect due to inscrutable evidence, if the datasets and algorithms used are not transparent.

Illustrative Example 3: A manufacturing organization would like to create a voice recognition system for automated robot task manipulation by shopfloor personnel. The sample that is used to train the voice recognition system will need to incorporate regional accents and have a balanced gender distribution. Otherwise, female employees or employees with regional accents may not be able to use the voice recognition system, facing an unfair disadvantage compared to other employees.
Risk: Biased training data leading to unfair outcomes.

Illustrative Example 4: An automated supplier performance monitoring system is being set up to rate suppliers of a large organization that produces engineering assets with a lifecycle of 20–25 years. Most data is collected from the organization's ERP system. However, the ERP system was implemented five years ago, and the collected dataset thus does not feature suppliers that have produced previous versions of the model, and this older data remains dispersed in individual spreadsheets. Some of the data is not accessible, as procurement officers who have developed a filing system have long retired. Thus, the performance monitoring system does not contain previous data on all suppliers, resulting in bias towards newer providers.
Risk: Missing training data leading to unfair outcomes.

Illustrative Example 5: A manufacturing organization introduces a task monitoring system on the shopfloor, for compliance certification and root cause analysis of any failures that occur during production. The system will monitor workers' body movements during manually intensive processes, in a bid to certify that the correct process steps were followed in the right order. The data collected to train the algorithm inadvertently contained samples from male workers, whose body shape and size are different, on average, of female workers. When the system is deployed on the shopfloor, jobs undertaken by female workers are frequently flagged up as incorrect, despite it being the contrary.
Risk: Biased training data leading to unfair outcomes.

Illustrative Example 6: A composites producer wants to develop a worker performance evaluation system for manually intensive production processes. The data collected to train the algorithm initially contains the gender of the worker. Realizing this feature could bias the dataset, as there are much fewer women operators, the company removes the gender variable from the model. However, the analysts do not realize that the dataset contains another variable that is correlated with gender, which is the shift identifier. The female workforce tends to prefer day shifts, due to caring responsibilities.
Risk: Biased training data leading to unfair outcomes.

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

Illustrative Example 7: A consulting company wishes to estimate carbon emissions during a set of production processes. As carbon accounting is a manually driven, complex process, the consulting company would like to automate the estimation by inferring carbon emissions of companies from companies that have already reported their metrics. It does so by creating a similarity measure, which takes into account company size, the sector and location it operates in, and typical production output. What is not known, however, is that the self-reported carbon emissions are incorrectly calculated in the first place, leading other companies, that have been found to be similar, to be adversely impacted by wrong estimates. Here, creating predictions from predicted data, the company has confounded multiple uncertainties, yielding uninformative scores.
Risk: Wrong data and aggregated uncertainty leading to unfair outcomes, inscrutable evidence.

The above examples pertain to challenges that may arise from biased data as well as issues with propagating uncertainties. We call for further research on the following questions:

- **RQ 3:** What sources of bias do manufacturing datasets and collection processes suffer from and how can they be identified and mitigated with minimal compromise on performance?
- **RQ 4:** How can informative, unbiased datasets be obtained from the shopfloor in contexts where humans are involved?
- **RQ 5:** How can workers who are unskilled in AI check for algorithmic bias and fairness?
- **RQ 6:** How can we ensure multiple sources of uncertainty in the AI supply chain are not propagated and amplified?

4.1.2. Data augmentation and pre-processing

Data augmentation and pre-processing refer to any techniques that enhance the size and quality of datasets involved in AI approaches, particularly in the case of data-intensive approaches within ML and deep learning. This may indicatively include labeling processes to convert an unlabeled data set to a labeled one, oversampling and undersampling techniques to address data imbalance, data cleaning, and feature engineering. We look at challenges related to each of these in this section. While data augmentation and pre-processing are central to machine learning workflows, analogous concerns arise in symbolic AI approaches. In knowledge-based systems, this stage may involve formalizing domain knowledge, validating rule sets, or refining ontologies. In both paradigms, early-stage decisions about data or knowledge representation directly impact the trustworthiness, interpretability, and performance of the resulting system.

Labeling. A label, in the context of supervised ML, is the value of an outcome variable that the ML model being developed will predict from the input data. For example, suppose one would like to predict the quality of a product from process parameters. We would need to obtain a set of data samples on products that were produced, which then relate process parameters to a quality indicator. In many real-world manufacturing scenarios, such indicators may not be easily available. Hence, one would need to label data manually. The volume of data might be too large to manually handle, meaning a sampling approach must be taken, where one must ensure an appropriate amount of samples are used, with appropriate variety. Alternatively, casting the problem as an unsupervised classification problem might be helpful. In addition to this, in many applications, the label is subject to the operator's expertise, environmental conditions, or time of measurement (van Giffen et al., 2022). Large-scale labelling is often outsourced, potentially to non-experts or to those without sufficient domain knowledge or who will not be fully aware of the intended application context, which is hard to monitor and validate and can lead to various issues downstream (Cobbe et al., 2023).

In other cases, the label of an outcome is uncertain. For instance, the expert's opinion on whether an outcome is favorable or fits into a given category might be debatable, in which case multiple experts must

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

Illustrative Example 8: A powder metallurgy company producing automotive parts wishes to create a quality prediction algorithm that will relate powder packing and subsequent sintering steps to resulting dimensional variance as a proxy for product quality. The production takes place in batch sizes of 1000. Due to the manual effort involved in generating the labeled data, the company opts for sampling 5 products at each process step. However, the samples during the process cannot be tracked individually; hence, at each production step, different samples are taken. Because of a loss of traceability, input parameters cannot be related to the resulting quality proxy. Further, the sample size is not sufficiently representative of the variety of input parameters. The company, therefore, opts to use an unsupervised learning algorithm to alleviate the labelling problem. Here, an autoencoder approach was deployed to detect outliers as dimensional anomalies, yielding a better proxy for quality prediction.
Risk: Small sample size, and lack of labels could result in misguided evidence.

be consulted, and a label should be agreed upon. Such instances are often the case when the ML task is a natural language processing on human-generated text or speech, such as in the illustrative example that follows.

Illustrative Example 9: A study was conducted to automatically extract supply chain maps from online text data. A natural language processing methodology was used to identify companies and determine supply relationships between them. Due to the size of the dataset, Amazon Mechanical Turk was used, which is a crowdsourcing approach where human labellers are paid to annotate text data. The complexity of the task was such that labellers frequently did not agree whether a given sentence constituted a supply relationship. Thus, multiple expert labellers were tasked in accordance with increased sentence complexity, and majority voting methods were used to determine the likelihood of a true label.
Risk: Uncertain labels leading to a wrong model, which can result in unfair outcomes, inconclusive, inscrutable, or misguided evidence.

In this context, majority voting is typically used. Alternatively, experts may be given different weights depending upon experience. Du and Ling (2010) suggest that these approaches simplify the problem by assuming uniformly distributed noise over the sample space, which fails to precisely reflect the human behavior in real-world situations. For example, when a human is highly confident in labeling outcomes, they are naturally less likely to provide incorrect answers, whereas when such confidence is low, the noise would be more likely to be introduced. They propose "noisy label oracles"—an active learning algorithm to simultaneously explore the unlabeled data and exploit the labeled data. Peyre et al. (2017) propose weak annotations for unusual or rare labels. However, imprecise labels can lead to a loss in quality of the model, making them unusable in safety-critical manufacturing contexts.

Data imbalance. In many manufacturing scenarios, the target of prediction is a rare event or outcome, creating a data imbalance issue. In the context of classification, data imbalance refers to cases where the positive class, i.e. the event being predicted, is by definition much rarer than the negative class i.e. an event not occurring. This may result in increased false positive rates because the biggest source of training data for the algorithm is in the majority, the negative class although it is the positive class that is the main target of the predictive process. In the context of manufacturing, this bias in predictive models results in the majority of faults going unnoticed (Fathy et al., 2020). Data augmentation approaches that help with data

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

imbalance include under- or oversampling, or generative methods where synthetic data is generated to counterbalance the minority class. Alternatively, algorithmic approaches, also called "cost-sensitive learning," can be used. Here, an artificial bias is implemented in the existing classification process through a cost function that amplifies the penalty value for misclassifying minority samples.

Although class imbalance is frequently due to the nature of the data itself, at times the labelling process itself could be to blame, such as in the following illustrative example.

Illustrative Example 10: An engineering company wants to predict the root cause of delays during production. They designed an interface attached to each workstation, which asked operators to indicate the reason behind disruptions when they took place. The root causes included operator error, tool unavailability, machine breakdown, and random stoppage. This approach resulted in severe class imbalance, as operators almost never selected operator error, and machine breakdowns were a rare occurrence. Operators perceived data collection on disruptions as time-consuming and stated that the default cause would often be a random stoppage. Had the company simply used this labeled dataset, it would have misdiagnosed the reasons for delays, perhaps increasing tool buffers.
Risk: Wrong training labels leading to wrong or sub-standard model, resulting in unintended consequences.

For the above use cases, the research questions raised in Section 4.1.1 pertaining to obtaining informative and unbiased datasets is relevant. Additionally, we ask:

- **RQ 7:** What are the best practices to tackle imbalanced datasets in a manufacturing and supply chain context?
- **RQ 8:** Which methods are most appropriate for managing uncertain labels in which manufacturing contexts?
- **RQ 9:** How can we make sure any automated labelling done to alleviate error can still leverage the operator's expertise?

Data cleaning. Data pre-processing is commonly needed to ensure datasets are meeting requirements of the AI algorithms they are fed into. The most significant part of pre-processing, and where a large amount of effort is arguably devoted, is data cleaning (Géron, 2019). This may involve identifying imputation of missing values, transformation of data into a form that is applicable, and, if necessary, reduction of the size of the dataset. Detection and removal of errors and decisions on whether a data point constitutes an error or an outlier, is an important aspect of the data cleaning process, which, if not done properly, can result in similar issues to biased data collection, potentially yielding inconclusive evidence, unfair outcomes, and inscrutable evidence.

Illustrative Example 11: A company uses goods-receipt data from one of its warehouses to predict when orders will arrive, so as to optimize stock. Upon inspection, the data analytics team finds that the prediction system flags items due on Friday as three days late. Further analysis shows that the items are not late indeed, but often do not get logged onto the purchasing system until the following Monday because of reduced numbers of warehouse workers on Fridays. Had this issue not been noticed, suppliers that deliver on a Friday would have been disadvantaged, as they would be categorized as low-performing suppliers.
Risk: Incorrect data leading to misguided and/or inscrutable evidence, unfair outcomes.

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

Illustrative Example 12: A train manufacturer would like to use samples with metal particulates in engine oil as a predictive feature for their prognostics algorithm, which will be used for planning maintenance. The data analytics team find out that the metal particulates for a particular train are not increasing with wear and tear as they should, but at times decreasing instead. A member of the team is sent to follow the train in operation, who finds out that the engine has an oil leak, which is being topped up as it moves across the route, making the underlying data irrelevant to the prediction.
Risk: Incorrect data leading to unsafe model operation, misguided and/or inscrutable evidence.

Illustrative Example 13: An analyst would like to predict product dimensions resulting from a 3D-printing process by using historical data. The analyst opts for a classification approach using a neural network but does not standardize the input features, resulting in non-activation of neurons, and the result does not offer better performance than random choice.
Risk: Lack of ML skills leading to substandard model and inscrutable evidence.

Feature engineering. As part of the augmentation and pre-processing phases, it is also common to explore whether additions to variables, rather than samples, are appropriate. In these cases, feature engineering is conducted, which involves creating new predictor variables from the original dataset to improve prediction capability. Successful feature engineering is highly dependent on domain knowledge. Experts need to agree on quantifiable hypotheses that can improve the prediction that can be extracted from the available data. Once features are created, it is important to identify whether those hypotheses were correct or not, which is influenced by the model selection as follows.

Illustrative Example 14: An example of the prediction of order delays from goods receipt data illustrates successful feature engineering. Here, existing features are used to predict whether an order would be delayed. These include supplier identification, locations the product is coming from and traveling to, the product name, contractual delivery duration, and the time an order was given. In addition, one of the hypotheses put forward by the procurement team is that if a supplier is more "agile," its orders would be less likely to be delayed. When prompted about how agility could be quantified from the existing dataset, the team designed a feature that analyzes how frequently a supplier was responding positively to changing demand patterns. This feature affirmed the initial hypotheses and led to a better predictive outcome.
Risk: Unexploited features, leading to a substandard model and inscrutable evidence.

The above cases and discussion highlight a need to ensure domain knowledge is incorporated in the data collection effort for the prevention of errors. However, doing so should not introduce new bias as experts impose their own values and priorities into the context. This leads to the following research questions:

- **RQ 10**: What are the best methods to ensure domain knowledge is fed into AI projects in a non-discriminatory way?
- **RQ 11**: How can we use this domain knowledge to automate model development, ensuring quality standards are met?
- **RQ 12**: What digital skills should manufacturing workers be equipped with so that a good synergy between the manufacturing expert and the data expert can be achieved?

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

4.2. Model creation

Following the data management part of the AI lifecycle, an AI model is created either through a model learning process, in the case of ML, or through knowledge engineering, in the case of knowledge-based AI. In this section, we focus primarily on challenges affecting model learning. This is because knowledge engineering is a human-centred process; as such, trustworthiness is less likely to be compromised as a result of the model creation process itself and is more a reflection of the trustworthiness of the human experts involved.

4.2.1. Model selection

Model selection refers to selecting the type of model that will be learned. The selected ML model influences its interpretability. Failure to obtain adequate interpretability may result in inconclusive evidence and inscrutable evidence. When interpretability is prioritized, the ability to interpret the output of a model plays a critical role in model selection, which then has to be balanced with computational cost as well as performance considerations. For example, decision trees (DT), which are a basic and effective ML algorithm, are widely used in practice (Baryannis et al., 2019). Both Baryannis et al. (2019) and Hansson et al. (2016) describe several cases in manufacturing that adopt DT because of their interpretability, ranging from supply chain risk prediction to steel production and continuous processing. Baryannis et al. (2019) caution against performance loss when opting to use simpler models, and casts the model selection challenge as a trade-off between performance and interpretability. It is thus important to consider multiple dimensions when it comes to model evaluation, ranging from common metrics such as accuracy to fairness and interpretability. The relationship between these dimensions may not necessarily be linear, and trade-offs may not be obvious. Interpretability may not necessarily mean poor outcomes, where a more interpretable or fair model might yield a small performance loss.

Illustrative Example 15: The purchasing department of a manufacturing company wants to predict quotes to be received from suppliers in advance, which could then be used to detect pricing anomalies. The company collects a dataset of all of the previously ordered products, which ranges from products that are highly complex to produce to simpler parts. A number of hypotheses are put forward by the lead procurement officer. Among them are the effects of multi-sourcing and legacy parts. The procurement office thinks that multi-sourcing caused a deterioration in the significance of individual supplier relations, hence parts that are bought from more than one supplier would be more expensive. As legacy parts are being discontinued, suppliers would think that the procurement office would be "locked in" to the relationship, unable to change suppliers, hence the price would typically be higher. A price prediction model is built using a Gradient Boosted Regressor, which shows key disagreements with the hypotheses. Neither legacy parts nor multi-sourcing are significant factors, but the main factor is the supplier being asked for the quote. Further analysis divides the dataset into price buckets and produces multiple models. Here, the importance of features shifts: for more expensive parts, it is the part complexity that affects price, whereas for simpler parts, price is determined by supplier discretion. As the company's understanding of what drives prices grows, they are able to better focus its purchasing strategy.
Risk: Lack of model interpretability leading to inscrutable and/or inconclusive evidence.

It is also important to note that interpretability may differ with differing model setups even when using the same learning algorithm. For example, features that were ranked to be important in a model constructed from a dataset may not be the same features when the dataset is filtered. An illustrative example is given below.

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

An additional reason behind the selection of simpler models is the lack of adequate computational resources, especially in resource-constrained environments, where energy, memory consumption, and data transmission are limited. For example, in offshore environments or agricultural production, data transmission is limited; hence, advanced techniques, such as deep learning, are not yet considered for practical deployment, despite being able to handle high-dimensional data. Here, the use of simpler models may lead to reduced performance and hinder trust in AI algorithms. There is some work done on the development of "white-box models" from "black-box models" (Alaa and van der Schaar, 2019), by relying on symbolic (knowledge-based) AI models. However, more work needs to be done to improve the accuracy trade-off when extracting these symbolic white box models, particularly in manufacturing environments where the margin of accuracy is fundamental.

In summary, the model selection phase involves two key issues: (i) interpretability versus model performance, and (ii) the consideration of computational resources and the environmental footprint of model training. Hence, we ask:

- **RQ 13:** How can we build rigorous processes to ensure the resulting outputs from ML models are explainable and interpretable in manufacturing scenarios?
- **RQ 14:** How can practitioners be effectively guided towards selecting the range of considerations to be prioritized for building ML models in differing contexts?

4.2.2. Model training and dataset concerns

This phase involves training the chosen model with the collected and processed dataset to learn patterns or representations of the data such that the model can then be used to cluster or classify newly observed inputs into groups, create continuous valued estimations about a new observation, or decide on a new action to take based on an expected value. For knowledge-based or symbolic AI systems, training corresponds to the process of rule formalization or ontology construction, in order to encode expert knowledge. While complex, this process does not generally require significant resources. In contrast, most ML models require hyperparameters to be optimized during the training process, such as the depth of a decision tree, the number of hidden layers in a neural network, or the number of neighbours in a K-nearest neighbors classifier. Finding the optimal settings of these hyperparameters requires multiple training rounds to be run. In the worst case, the size of the hyperparameter optimization search space grows exponentially. Thus, as mentioned earlier, one of the biggest concerns with the model training stage is the economic cost associated with carrying out the training procedure due to the computational resources required. Strubell et al. (2020) also raise an additional, growing concern around the environmental cost of training, showing that a full training cycle on a neural network could emit carbon dioxide comparable to carbon emissions of four average cars in their whole lifetime. This is especially the case when large-scale language models are concerned. We depict carbon emissions of model training as a case of unintended consequences of ML-based AI, as companies are often unaware of its environmental cost.

- **RQ 15:** How can companies make informed decisions that consider not only the cost of building ML models, but also their carbon footprint?

Another concern regarding parameter tuning is the skillset required. If the AI team is not well rehearsed in parameter tuning, the outcomes might be sub-optimal. Alternatively, automated ML (AutoML) toolkits (including online services) could be used, which automate various stages of the ML, including hyperparameter tuning. However, the criticism with these is that they make the process of developing ML models even more opaque, resulting in inscrutable evidence. Krauß et al. (2020) describe a use case where AutoML was pursued to predict out-of-specification products and concluded that data science expertise is necessary and cannot be completely replaced by an AutoML system. For example, data integration, the handling of instability of training, and inefficient management of the hardware

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

resources were a challenge. Overfitting was encountered, which was overcome by manual intervention. Similarly, automating various ML processes can potentially optimize for certain (often functional) aspects, while potentially ignoring aspects that might be more broadly relevant, be they around issues of bias, transparency, privacy, and so on (Lewicki et al., 2023; Sun et al., 2023). For example, it has been shown that AutoML platforms might select for a user a model with the highest accuracy but which is highly biased, at the expense of models with slightly less accuracy but with much reduced levels of bias (Lewicki et al., 2023). A broader point is that AutoML tooling serves to operate generically to support anyone seeking to build a model, and therefore generally will not, nor cannot, account for all the issues for the potential contexts in which the models that are automatically built will be deployed (Lewicki et al., 2023).

Illustrative Example 16: Rising energy prices increasingly necessitate more careful budgeting for production facilities. A machine tool producer would like to create a predictor for energy consumption at their factory using sensor-based data as well as features such as production schedules and machine attributes. The company does not have a data science team and cannot afford to hire a specialist consultant. They therefore opt to use an AutoML library to automate part of the ML pipeline. The results seem promising, with over 95% accuracy, and the tool is deployed to budget for energy bills. After a few months, it is noticed by the accounting team that the tool vastly underestimated the energy consumption, as a result of overfitting to existing datasets and inadequate training on changing production schedules.
Risk: Lack of ML skills leading to wrong models which can cause misguided evidence. Such models may be inscrutable.

With the availability of numerous open source AI libraries, the reuse of data and ML models might become increasingly commonplace, which is helpful for saving time and effort but comes with no security guarantees (Gu et al., 2017). Code reuse, however, also creates potential security issues. One of the main privacy issues concerns the preservation and leakage of the datasets collected by companies, for example, through adversarial attacks that allow data reconstruction (Shokri et al., 2017) or data poisoning (Terziyan et al., 2018). For example, Yampolskiy et al. (2021) illustrated how a self-learning, Internet of Things (IoT) connected 3D printer can be corrupted by the injection of a small number of wrong labels. Researchers have developed a multitude of technical frameworks to preserve privacy during the training cycle, including explicit corruption of the data with differential privacy (Dwork, 2006), encrypted training (Gentry, 2009), and federated learning that distributes training across personal devices to preserve privacy (Zheng et al., 2023). These are relevant to manufacturing, especially in federated learning use cases and wider agent-based distributed learning, where datasets from multiple entities (Yong and Brintrup, 2020) or organizations are used to create common predictive models, such as supply chain disruption prediction (Zheng et al., 2023), industrial asset management (Farahani and Monsefi, 2023) and prognostics (Dhada et al., 2020).

Using an interview-based methodology, Kumar et al. (2020) found that industry practitioners were not equipped with the tools to protect, detect, and respond to attacks on their ML systems. The interviews revealed that security analysts either diverted responsibility to the company ML service is bought from, or expected algorithms available in commonly available platforms such as Keras, TensorFlow, or PyTorch to be inherently secure against adversarial manipulations—which is not the case. The authors recommend more research to be undertaken in areas such as automated testing against adversarial attacks, threat modeling, containerization, and rigorous forensics.

While manufacturing cybersecurity is out of the scope of this review, it is important to note that model poisoning attempts can be made by adversaries that are outside the organization, who may gain

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

access to ML models via IoT systems, widely deployed in industry. IBM's recent X-Force Threat Intelligence Index found that the manufacturing sector was the most attacked by ransomware, accounting for 23% of reports. Manufacturers are especially vulnerable to the algorithmic supply chain, as cyber-physical systems that are deployed are increasing. For example, industrial robots have grown from 54,000 supplied in 2010 to 121,000 in 2015, many including IoT components, which pose another point of entry to industrial information systems. As robotic systems are difficult to update and deploy virus checks on due to costly downtimes, they may make AI systems vulnerable to attacks. Thus, efforts to prevent data reconstruction and model poisoning should include cybersecurity checks, as illustrated in the example that follows.

Illustrative Example 17: A group of attackers identifies a cybersecurity weakness in a manufacturer's newly installed vision recognition system used to detect objects on the work-in-progress buffers. The system is used to automatically update the company's inventory management system by keeping track of production quantities. The adversaries implement a data poisoning attack, injecting bad data into the system, causing it to misclassify objects. This is only noticed when the inventory management system gives a number of automated orders to the company's suppliers for presumably out-of-stock items, which were, in fact, in stock.
Risk: Adversarial attacks designed for any harmful outcomes, such as unsafe operation, unethical or biased models.

Adversarial attacks on models are thus a real concern, especially in infrequently updated cyber-physical manufacturing systems. We ask:

- **RQ 16:** What are the ways in which adversarial ML attacks can take place in manufacturing, and how can they be prevented?

4.3. Ethical governance and Interoperability

A growing concern is the privacy of personal data used to train AI, primarily ML-based ones. In a manufacturing scenario, this may involve end-user (customer) data, as well as supplier, business relationship data, and data from employees. The ethical implications of violating worker privacy are a growing concern that crucially needs more attention in the context of manufacturing. Surveillance mechanisms deployed on the factory floor are a prime example (De Cremer and Stollberger, 2022). Although manufacturing-specific surveys have not yet been conducted, in 2017, a global survey found that over 69% of companies with at least 10,000 employees have an HR analytics team that uses automated technologies to hire, reward, and monitor employees.

In many cases, shopfloor worker monitoring may have been deployed with valid and ethical intentions. These may include, for example, ensuring Personal Protection Equipment (PPE) has been worn correctly, identifying hazards, aiding novice workers with suggestions on how to complete a difficult task, quality certification of products that necessitate an evidence trail that processing steps were performed adequately. However, the same technology can be used in ways that are ultimately detrimental to the well-being of workers, erode human-centered values, and jeopardize individual rights to self-determination.

Unsurprisingly, there are limited cases that have been brought to light, and even fewer academic studies. One of the high-profile cases has been reported by the Open Markets Institute, an advocacy group focusing on technology company monopolies (Hanley and Hubbard, 2020). They found that Amazon uses a combination of tracking software, item scanners, wristbands, thermal cameras, security cameras, and recorded footage to monitor the activities of warehouse workers. Whistle-

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

blowers suggested that workers need to wear an item scanning machine (scan gun) which detects "idle time." The scan gun alerted a manager if workers spent over the maximum allowance of 18 minutes of idle time per shift. Idle time included bathroom breaks, getting water, or walking slower, and thus could be easily exceeded. The algorithm that powered the scan gun would classify idle time based on expected levels of motion. Two other cases from Amazon included recognizing when a forklift driver has been yawning, which the drivers saw as an invasion of their privacy; and the use of employee personal data in conjunction with shopfloor worker monitoring data to prevent unionization. In cases such as worker monitoring, there is an inherent power imbalance between the employer and employees, making it hard for workers to question data being gathered about them and algorithms used to analyse their data.

Remote working during the Covid-19 pandemic has increased reports of privacy invasion, for example, by software that detects worker productivity through monitoring keyboard strokes. Other commercially available software allows company managers to map company social networks by using email metadata and detecting employee "sentiment" through email conversations, and even predicting when an employee is showing signs of frustration and may want to leave the company. The market for HR Analytics software, which includes manufacturing worker surveillance, is projected to reach USD 11 billion by 2031.

Although worker monitoring itself is not a new concept, in the case of AI the fear is that monitoring can be scaled up by including multiple, often objectionable and private data sources (*transformative outcomes*), and inference is automated without any real insight to the algorithmic decision process or data itself (*inscrutable and inconclusive evidence*), potentially resulting in discriminatory practices and undue pressure on employees, effecting their well-being (*unfair outcomes*). The power imbalance between workers and managers means that workers often have no say or are hesitant to say whether and how such technology should be adopted.

In response, several countries are proposing regulations to prevent ethical issues arising from AI, but more research is needed to inform regulators of the potential consequences of the misuse of algorithms to benefit commercial interests in the manufacturing workplace, and ensure that regulations are effective. The White House Office of Science and Technology issued the Blueprint for an AI Bill of Rights in October 2022 (Office of Science and Technology Policy, 2022), with the aim of protecting civil rights and democratic values in the development and use of automated systems. The Bill of Rights highlights the need for data privacy in the employment context. Again, in the United States (and one province in Canada), three states have implemented laws that require employers to notify employees of electronic monitoring, including AI-powered technologies, by providing notice to employees whose phone calls, emails, or internet usage will be monitored. In New York, automated decision tools that replace or assist in hiring or promotion decision-making must undergo annual bias audits. Companies must make the audit results publicly available and offer an alternative selection process for employees who do not want to be reviewed by such tools.

In Europe, both Norwegian and Portuguese Data Protection Authorities outlawed the practice of remote worker monitoring. The European Union is drafting an Artificial Intelligence Act (EU AI Act [European Commission, 2021]) to regulate AI, in which "employment, management of workers, and access to self-employment" are considered high risk and will be heavily regulated. The EU's General Data Protection Regulation (GDPR) limited the use of AI in employment, explicitly stating that employees should not be subject to decisions "based solely on automated processing." The UK has published an AI regulation white paper (Department for Science, Innovation and Technology, 2023). The white paper's focus is on coordinating existing regulators such as the Competition and Markets Authority and Health and Safety Executive, but it does not propose any regulatory power. Critics raised that the UK's approach has significant gaps, which could leave harms unaddressed, relative to the urgency and scale of the challenges AI brings (Hern, 2023).

The White House Office of Science and Technology issued the Blueprint for an AI Bill of Rights in October 2022, with the aim of protecting civil rights and democratic values in the development and use of

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

automated systems. The Bill of Rights highlights the need for data privacy in the employment context. Again, in the United States (and one province in Canada), three states have implemented laws that require employers to notify employees of electronic monitoring, including AI-powered technologies, by providing notice to employees whose phone calls, emails, or internet usage will be monitored. In New York, automated decision tools that replace or assist in hiring or promotion decision-making must undergo annual bias audits. Companies must make the audit results publicly available and offer an alternative selection process for employees who do not want to be reviewed by such tools. Both Norwegian and Portuguese Data Protection Authorities outlawed the practice of remote worker monitoring. In the EU AI Act, "employment, management of workers, and access to self-employment" are considered high-risk and are heavily regulated. The EU's General Data Protection Regulation (GDPR) limited the use of AI in employment, explicitly stating that employees should not be subject to decisions "based solely on automated processing." The UK has published an AI regulation whitepaper (Department for Science, Innovation and Technology, 2023). The whitepaper's focus is on coordinating existing regulators, such as the Competition and Markets Authority and Health and Safety Executive, but does not propose any regulatory power. Critics raised that the UK's approach has significant gaps, which could leave harms unaddressed, relative to the urgency and scale of the challenges AI brings (Hern, 2023).

While the regulatory debate is encouraging, we note that the effectiveness of these regulatory initiatives in the manufacturing and supply chain sector remains to be seen. At the moment, these frameworks are not enforceable, and progress is not fast enough to keep up with the fast pace of AI research developments. Although multiple high-level frameworks have been created, as summarized in Section 3, there is a lack of practical use cases and guidance on how these frameworks can be incorporated into daily business practice. This is especially true in the field of manufacturing.

At the time of writing, ISO, IEC, and BSI standards are being discussed and proposed (https://aistandardshub.org/) (e.g., ISO/IEC TR 24028:2020). However, at the moment, there is no consensus on the adoption of trustworthy AI standards. Should companies wish to develop surveillance mechanisms on their workers to prevent, for example, unionization, they are free to do so. Participation in industry standards may play an important role in designing effective regulatory frameworks. The emergent interplay between standards development and regulatory approaches will be a decisive factor, and multiple, clashing standards and regulations might stifle progress in the area. Regulatory enforcement of standards may be criticized for stifling innovation, whereas too laissez-faire an approach may yield unintended consequences, as discussed in this section. Non-governmental organizations such as the Algorithmic Justice League have already helped scrutinize and remove bias from a number of facial recognition algorithms used by the police force in the United States. Similar approaches can be taken in manufacturing. Researchers also suggested that independent audit firms could develop reviewing strategies for AI projects and make recommendations to their client companies about what improvements to make. The insurance industry could also help guarantee trustworthiness by specifying requirements for underwriting AI systems in manufacturing.

We therefore call for more research in understanding how manufacturing organizations exploit AI technology in ways that can breach human privacy and well-being, and what mitigation mechanisms and guidance can be designed to prevent such breaches:

- **RQ 17:** Should we build a manufacturing sector-specific code of conduct that interprets and adapts existing legal instruments pertaining to the use of AI?
- **RQ 18:** How will we ensure interoperability between the various AI regulatory frameworks and standards that are currently being developed in different regions?
- **RQ 19:** How will multinational manufacturing organizations adopt differing standards in their supply chains?

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

Illustrative Example 18: A vision recognition algorithm has been designed for a factory shopfloor during COVID-19 to ensure workers follow social distancing rules and wear PPE. The system would warn employees when distancing rules were not met. The system was designed to be private and would hold no personally identifying information, only using generic object recognition. However, leaked documents and subsequent media interviews with workers have suggested that the system was combined with personal data, such as the number of complaints raised, and used for an additional purpose: prevention of unionization. The company used additional datasets on worker background and personal information to estimate a risk score for unionization. Based on the score, high-risk individuals would be warned to keep a distance from certain individuals or reallocated to different shifts.
Risk: Unethical surveillance or misappropriation of data leading to privacy violation, unfair and/or transformative outcomes.

Illustrative Example 19: A supply chain surveillance algorithm is deployed to help improve supply chain visibility at a company. The tool will improve an understanding of which geographic locations the company's upstream suppliers are concentrated on, so that risk mitigation measures can be taken. The tool predicts a link between one of the company's suppliers and an "anonymous firm." Although the firm name is anonymous, because of the additional information revealed, including geographic location and production, the company infers that the supplier has also been selling to their main competitor.
Risk: Unethical surveillance or misappropriation of data leading to privacy violation, unfair and/or transformative outcomes.

Brundage et al. (2020) identified institutional governance to be another key candidate for improving ethical practices. They suggest visible leadership commitment, including regular review board meetings, annual, publicly available responsible AI reports, and reward mechanisms for responsible AI practices, can be valuable in increasing incentives in organizations. Brundage et al. (2020) also suggest inter-institutional reporting mechanisms such as NASA's Aviation Safety Reporting System and the Food and Drug Administration's Adverse Event Reporting System, and Bugzilla as useful models for technical reporting.

As seen above, a large number of privacy and ethical challenges can arise during the model training phase. Data collection and model training are intertwined when it comes to trustworthiness. While in Section 4.1.1 we highlighted that data needs to be bias-free and obtained under consensus, and be able to be scrutinized by the owners and generators of data who might be unskilled in AI, in this section, we raise additional questions pertaining to personal data that is used to train ML models. These include:

- **RQ 20:** What is the definition of personal data in a manufacturing context? How can workers know how their data is used, and can they have a right to consent or decline the way their personal data is used?
- **RQ 21:** Should policymakers build mechanisms to ensure that data that was originally collected for its purpose remains its purpose in a manufacturing context?
- **RQ 22:** What robust institutional mechanisms can be put in place to empower employees to scrutinize AI models in an organization? What methods can be used by an unskilled workforce to check for algorithmic bias and fairness, as well as decision processes that impact them, which rely on AI?
- **RQ 23:** Counterarguments on surveillance activities on suppliers and workers highlight that surveillance has always been practiced, and the only difference AI brings is scale and accuracy. Does the manufacturing community, including generators of personal data, agree with this

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

statement? In a manufacturing context, can we achieve consensus on what types of data collection and algorithmic surveillance constitute fair and unfair outcomes?

- **RQ 24:** For unethical practices, how can the right balance between regulation and innovation be found? Should specific AI standards for manufacturing be built, and if so, should they be enforceable in different contexts?

4.4. Verification

Verification of AI models should include rigorous checks to ensure they are robust and reliable in satisfying functional and performance requirements. Here, robustness refers to the degree to which the developed model can function correctly in the presence of invalid inputs or varying environmental conditions, while reliability refers to the probability that the model performs required functions for the desired period of time without failure. Verified artificial intelligence has been defined as "AI-based systems that have strong, ideally provable, assurances of correctness with respect to mathematically-specified requirements" (Seshia et al., 2022).

In the case of knowledge-based models, verification is rooted in the strong mathematical logic-based foundations of such models and leverages extensive research and development efforts in model checking (Clarke et al., 2018) and automated theorem proving (Bibel, 2013), as well as formal specification languages (Baryannis and Plexousakis, 2013, 2014; Baryannis et al., 2017). For instance, ontology-based knowledge models can be verified through an array of established reasoning systems, such as HermiT (Glimm et al., 2014) or Pellet (Sirin et al., 2007), available through established, user-friendly tools, such as Protégé.

In contrast, verification of neural network-based ML models is a hard problem due to their black-box nature, making approaches such as the aforementioned model checking or theorem proving, or even source code reviews, not applicable (Salay et al., 2017). One of the issues is the large size of the state-space, making the design of test cases difficult. Reinforcement learning (RL) approaches especially suffer from large state spaces, as the decision space is non-deterministic and the system might be continuously learning, meaning that over time, there may be several output signals for each input signal.

Automated test case generation (Clark et al., 2014), transfer learning and synthetic data generation (Borg et al., 2018) have been proposed as potential solutions. El Mhamdi et al. (2017) suggested that the robustness of a deep neural network could be evaluated by focusing on individual neurons as units of failure. Continuous monitoring of model input (elaborated further in Section 4.5), and integrating verification processes across the whole development cycle rather than at the end (e.g., by experimenting how output varies in the state-space with different model architectures) have been proposed as best practice (Taylor, 2006). Adler et al. (2016) proposed coding a "safety-cage," where the model execution is turned off with increased uncertainty and swapped with a deterministic model track. Although these suggestions stem from the field of autonomous systems, notably autonomous vehicles, they are worth noting and re-interpretation within the context of manufacturing is worth exploring.

Test-based verification usually involves the setting up of a simulation-based test environment, which is typically safer, cheaper, and faster to run. However, as with any simulation-based methodology, conclusions derived are dependent and constrained by the assumptions made by the simulation designer. Even small discrepancies between the simulation environment and the real world can cause dramatically different outcomes, exemplified by high-profile cases in the field of autonomous vehicles (Dulac-Arnold et al., 2021). Focusing on RL, the authors highlight several challenges with the transfer of RL algorithms from simulation-based training to real-life environments. These include limited sample size, delays in task rewards, constraints, unexpected, stochastic changes in the environment, and multiple objective functions. Stochasticity means that agents are not guaranteed not to explore unsafe conditions, which may have unintended consequences, unless these are thought by the designer in advance and coded into the reward function, which is often infeasible for the designer to capture exactly what they want an agent to do. Consider a robot tasked to pick up items in a warehouse from delivery zones and place them in

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

designated locations. The algorithm designer may simply code a reward function to maximize the number of items picked and placed. In the simulation, the robotic agent works in a fairly constrained, stable environment. In reality, the layout of the warehouse may change, with moving obstacles that may include human workers. If the agent has not been trained to avoid moving obstacles, and its reward is based on number of tasks completed, it could explore taking unsafe shortcuts (called reward-hacking). While RL shows much promise as a learning paradigm, its implementations in real-world settings remain very limited (Z. Wang and Hong, 2020). As RL is starting to be popularized in manufacturing robotics (Oliff et al., 2020), condition-based maintenance (Yousefi et al., 2020), vehicle-routing (Mak et al., 2021), and inventory control (Kosasih and Brintrup, 2022; Z. Wang and Hong, 2020), it is worth exploring these challenges in real-world manufacturing environments.

Illustrative Example 20: A company would like to implement autonomous cleaning robots in its warehouse to speed up operations after a shift ends. The company has bought ML-as-a-service from a well-known AI solution provider. The provider has multiple success stories with warehousing, giving confidence to the purchasing company in its credentials. In a bid to speed up the training process, they use transfer learning, which involves extrapolating a reward function for the new warehouse environment based on reward functions from other similar cleaning robot algorithms they have developed. The provider also sets up a period of observation in the warehouse to ensure this approach would work in the new setting. For a period, the pilot seems successful. Following a change in the cleaning materials used, an incident happens, leading to a fire as the robot follows an unsafe shortcut where a chemical process is taking place, as it was not explicitly coded not to do so, which should have been the case after transfer learning.
Risk: Insufficient verification leading to unreliable, unsafe operation.

ML performance metrics should be carefully considered in the application context. While the accuracy of a classifier is a well-known and widely used metric, in safety-critical applications such as machine failure, other metrics may be more appropriate (Baryannis et al., 2019). For instance, recall may be more important, which measures the number of correctly identified positive classes over all classes that should have been identified as positive. In other cases, where incorrect identification of a class is costly, precision may be used—for example, when a production or quality delay is falsely predicted, resulting in inventory build-up. Other considerations may include quantifying the uncertainty of predictions, both in a classification and a regression context. Hence, performance metrics should be carefully designed and reflect contextual priorities. Additionally, performance checks should include checking for bias and fairness to overcome some of the issues relating to unfair outcomes, discussed in previous sections. Here inherent bias in the training data can be checked by comparing the ranges of features to the actual distribution of the feature in the real world across different data slices.

Illustrative Example 21: Consider a manufacturer developing a classifier for predicting supplier delays, which will then be used to optimize buffers. The manufacturer may choose to optimize the training cycle using precision or recall. Precision refers to the ratio of correctly predicted delayed orders over all delayed-order predictions, and recall refers to the ratio of correctly predicted delayed orders over the number of actual delayed orders. False classification of an on-time order could lead to unnecessary risk mitigation actions, such as building inventory buffers that might be costly. On the other hand, false classification of a delayed order as low risk could be more problematic, as the costs of dealing with an unexpected disruption could outweigh mitigation planning. The manufacturer may need to weigh these objectives, for example, using an F-measure, which allows one to combine these two objectives and weigh each one differently.
Risk: Incorrect model objectives leading to unintended consequences.

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

The main concerns at this phase include the safety and reliability of the developed models in noisy manufacturing contexts:

- **RQ 25:** How can established approaches in knowledge-based AI verification, such as model checking, be leveraged in the case of ML?
- **RQ 26:** How can test case generation methods developed be applied to manufacturing use cases?
- **RQ 27:** How can the stochasticity of manufacturing environments be captured in ML test environments in a meaningful way?

Further, during the training phase, performance metrics can have varying impact on safety, cost, and efficiency. We ask:

- **RQ 28:** How do different performance metrics affect outcomes in differing manufacturing contexts?
- **RQ 29:** How can insurance and/or legal coverage be ensured for continuously adopting AI models?

4.5. Model deployment

Model deployment refers to the operationalization of the model that has been built, by building the software infrastructure that is necessary to run it, and setting and following policies on model maintenance and updates. One of the major challenges in this stage is identifying when a model needs to be updated with new information, and to what extent older information should continue to be utilized. In the case of ML models, "concept drift" describes the situation where the feature distribution shifts over time due to a change in the underlying data-generating process. Concept drift means that the mapping between features and output no longer matches the new incoming data. Thus, as real-world contexts evolve and adapt to changes over time, the underlying datasets that are representative of the system should change. For example, a demand forecasting model used to predict demand for a fashion product should be retrained frequently. However, in other cases, the changes in the system may not be contextually obvious to the model owner, in which case input data should be continuously monitored.

The two main ways to deal with concept drift are to update the model incrementally or to retrain the whole model, which can be done periodically, based on predefined performance criteria (such as accuracy or F1 scores) or statistical approaches based on uncertainty quantification (such as Hoeffding bound). Concerns akin to concept drift also apply to symbolic systems, where rules may degrade in effectiveness as operational contexts shift, necessitating mechanisms for verification, updating, or human-in-the-loop validation.

Illustrative Example 22: A forging machine is equipped with a condition monitoring algorithm to estimate time to failure, based on the number and desired shape of parts that it handles. While this is initially successful in reducing unplanned downtime by correctly estimating service needs, over time the accuracy of predictions decreases. Upon inspection, it is found that the reason is that the material specification of the main batch produced by the machine has slightly changed along with the supplier of the raw material, creating higher loads on the machine.
Risk: Concept drift leading to misguided evidence.

Through this illustrative example, it is made clear that the model deployment phase is continuous. Challenges resulting from this involve finding the right frequency and method of model updating and detecting when a model is no longer applicable to the context it is deployed in:

- **RQ 30:** Which applications in manufacturing are prone to concept drift? Which drift detection methods are more informative in manufacturing?

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

- **RQ 31:** What are the best practice mechanisms to identify models used in manufacturing that are no longer useful and should be updated or decommissioned?

5. Cross-cutting considerations

In this section, we briefly discuss cross-cutting trustworthy AI challenges that manufacturing organizations face when considering the adoption of AI technology within their organizations or across their supply chains.

5.1. Affordability

A key issue in adopting AI technologies is cost, which not only includes time and effort spent across the development and deployment steps of AI, but also the cost of data access, storage, and post-deployment costs such as maintenance. Studies performed on the adoption of digital manufacturing technologies, which include AI, show that adoption is often contingent upon affordability. In the UK, for example, over 99% of businesses are small to medium enterprises (SMEs, 0 to 249 employees) with lower affordability, which might create a larger capability discrepancy in supply chains. It is worth noting that many SMEs are vital to the supply chains of larger organizations, hence the success of the manufacturing industry is intertwined. The manufacturing community needs to monitor and encourage SME adoption, and we propose the following research questions in this context:

- **RQ 32:** How can SMEs access the benefits of AI solutions?
- **RQ 33:** Does the affordability of AI impact trustworthy AI adversely?

5.2. Outsourcing AI as a service

While no current statistics exist on the extent to which AI is developed in-house versus bought as-a-service, both approaches are not without challenges for manufacturers, and outsourcing decisions are likely to depend on a number of factors including the specificity of development, longevity of its use, AI skills the company would like to retain, the degree of control a company wants to exert on the algorithmic approaches developed and external infrastructure dependencies. While classical theories such as Resource-Based View and Transaction Cost Economics may offer useful starting points for framing AI outsourcing decisions, additional considerations on the trustworthiness of algorithms may need to be factored in. As mentioned earlier, outsourcing AI may make it more difficult to investigate bias in data used to train algorithms, as well as affect algorithmic safety. Moreover, it might be tempting to outsource in an attempt to try and shift accountability and responsibilities elsewhere, which raises legal and broader governance considerations (Cobbe and Singh, 2021; Cobbe et al., 2023) We call for more research on AI outsourcing decisions:

- **RQ 34:** Are existing decision frameworks for outsourcing applicable for AI as a service in manufacturing and if not, how should they be extended?
- **RQ 35:** How can trustworthy AI be ensured when outsourcing AI in manufacturing?

5.3. Data compensation and monetization

Data is one of the most valuable assets of firms that fuel much of the digital economy today. Often, datasets are traded amongst companies by so-called data brokers. As a society, we often do not know how and where our data is used, and for what purpose, forming an active field of ethical and regulatory debate. In manufacturing scenarios, companies may use data not only from individuals (such as monitoring how a consumer uses their products, or workers producing them), but also from other companies (such as monitoring activities of their suppliers and competitors). At the moment, generators of these datasets are

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

typically not compensated. While this constitutes an ethical issue, in other cases data compensation may open up new markets and opportunities. For example, manufacturers may be interested in tapping into other companies' datasets that they lack. For example, if a manufacturer would like to implement prognostics for its machinery but does not have enough run-to-failure data, it may be able to appropriate datasets from another manufacturer using the same machine. The data owner may wish to monetize its datasets. The automotive industry has been discussing how customers can be compensated by sharing car usage data so they can design better (McKinsey and Company, 2020). Data compensation is a strongly debated field where no regulatory guidance currently exists. Researchers have proposed the use of blockchain for tracking how data is used and developing mechanisms to compensate owners of data (Maher et al., 2023). We call for more research on how manufacturers can monetize their datasets and also compensate data owners ethically and responsibly:

- **RQ 36:** Which types of manufacturing data can be monetized?
- **RQ 37:** Would monetizing data have an effect on AI trustworthiness in manufacturing and what are the associated risks and legal implications?

5.4. Scalability of trustworthy AI

Although some large manufacturers and supply chain businesses have made some advances in the incorporation of AI-based solutions into their decision-making and process control, for most companies (particularly SMEs), AI systems remain at the pilot stage (a situation often referred to as "pilot purgatory"). These pilot studies, as is the case with most AI solutions, implement models tailored to a specific use case and developed with data that manufacturers are typically not keen to share. However, as seen in the success of AI in other sectors such as finance or e-commerce, the secret to scalability and production-ready models is in sharing these models and data across businesses. Hence, to achieve the full potential of AI in the manufacturing value chain, it is recognized that models and data need to be transparent and usable, but at the same time secure to protect intellectual property and privacy (Davis et al., 2022). More research and development is needed to better understand the positive and negative effects of provenance on privacy and finding the best ways to develop and share models from aggregated data without losing sight of security issues:

- **RQ 38:** How can trustworthy AI solutions be made more scalable in manufacturing?

6. Conclusions

AI can be a significant driver in improving productivity in manufacturing and supply chains, but it can also be misused with unsafe, unethical practices. While in the Western World, there is progress towards a set of commonly agreed principles, there is significant confusion on what they mean in practical terms. As AI technology is moving rapidly in manufacturing, there is an urgent need to guide manufacturers on the risks that come with AI adoption and deployment, so that its benefits can be delivered safely and ethically.

In this paper, we have conducted a brief review of terminology in the field of trustworthy AI, after which we mapped potential risks that arise during the AI development and deployment lifecycle, using illustrative use cases. This thought exercise has shown that trustworthy AI risks may be present throughout the complete AI lifecycle, from data collection to post-deployment, as summarized in Table 2. We also highlighted a number of cross-cutting concerns that may be present throughout the entire AI development process. Doing so yielded 38 research questions aimed at guiding research into trustworthy AI in manufacturing and supply chains. In addition to guiding research, we hope that the mapping provided will help practitioners identify the types of risks they should pay attention to while they go through stages of AI development in their organization.

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

Table 2. *Summary of trustworthy AI challenges in manufacturing*

Process	Illustrative cases	Research questions	Issue	Trustworthy AI challenge	Trustworthy AI principles	Potential directions
Data collection	1	1	Missing data, wrong model	Inscrutable evidence	Valid and reliable	Develop common interoperable data schemas
Data collection	2	2	Missing or incorrect data	Inscrutable evidence, unsafe operation	Valid and reliable, safe, accountable and transparent	Leverage provenance mechanisms
Data collection	3–6	3–5	Biased or missing data	Unfair outcomes	Fair with harmful bias managed	Check data bias using fairness toolkits, establish ethics board
Data collection	7	6	Wrong data, aggregated uncertainty	Unfair outcomes, inscrutable evidence	Fair with harmful bias managed, explainable and interpretable	Check data validity
Data augmentation	8	7	Small sample size, lack of labels	Misguided evidence	Valid and reliable, responsible practice and use	Unsupervised learning, incorporate domain knowledge
Data augmentation	9	8	Uncertain labels	Unfair outcomes, inconclusive, misguided, inscrutable evidence	Valid and reliable, fair with harmful bias managed, explainable & interpretable	Revise data augmentation process, incorporate domain knowledge
Data augmentation	10	9	Wrong labels	Unintended consequences	Valid and reliable	Crowd-sourced labeling, Majority voting, Noisy oracles, Weak annotations
Data pre-processing	11–12	10–11	Incorrect data	Misguided, inscrutable evidence, unsafe operation	Safe, valid and reliable	Check data validity
Data pre-processing	13–14	12	Lack of ML skills	Inscrutable evidence	Valid and reliable, responsible practice and use	ML expert input

Continued

Table 2. *Continued*

Process	Illustrative cases	Research questions	Issue	Trustworthy AI challenge	Trustworthy AI principles	Potential directions
Model selection	15	13–14	Explainability	Inscrutable, inconclusive evidence	Explainable and interpretable, accountable and transparent	Explore explainable AI methods, Domain expert input
Model training	16	14	Lack of ML skills	Misguided evidence	Valid and reliable, safe, responsible practice and use	ML expert input
Model training	17	16	Adversarial attacks	Any harmful outcome, such as unsafe operation	Secure and resilient	Adopt cyber security measures
Model training	18–19	17–24	Unethical surveillance, Data misappropriation	Privacy, unfair, and transformative outcomes	Privacy-enhanced	Regulation, Adherence to standards, Establish an ethics board
Verification	20	25–27	Insufficient verification	Unreliable, unsafe operation	Valid and reliable, safe	Automated test case generation, Safety cages, Synthetic data, Transfer learning
Verification	21	28–29	Incorrect objectives	Unintended consequences	Valid and reliable, accountable and transparent, responsible practice and use	Use of appropriate performance metrics
Model deployment	22	30–31	Concept drift	Misguided evidence	Valid and reliable, secure and resilient, responsible practice and use	Adopt appropriate model monitoring and update processes

The research questions that have been raised can be classified into three main categories: risks pertaining to data collection and processing, algorithmic development and deployment, and organizational practice of AI in manufacturing.

6.1. Risks pertaining to data collection and processing

Incorrect, misrepresented, or historically biased data may all be present in manufacturing AI use cases. Our analysis showed that manufacturing is at risk of both specific instances of data collection bias (for example, shopfloor personnel labelling why errors occur in production), but also of discriminatory bias, such as the use of datasets with inherent bias on gender or ethnicity. Aggregation of bias is also a risk factor, as uncertain labels or measurements are used to train models. Another common problem in manufacturing stems from the target of predicting rare events such as machine failures or supply delays, which yield data imbalance. Organizations need to adopt skills to scrutinize bias in datasets and remove it. Researchers need to create both technical advances in identifying and removing bias, and correlate types of bias in datasets with the manufacturing scenarios that yield them. Simplifying bias exploration and empowering employees who are unskilled in AI with the tools to scrutinize datasets is another important gap that researchers may wish to focus on.

6.2. Risks pertaining to algorithmic development and deployment

Algorithmic risks are those that stem from inappropriate or insufficient design and use of AI algorithms. Researchers have highlighted the trade-offs between explainable and interpretable algorithms and performance, as well as noting the often large carbon footprint of model training. Future lines of research should consider the exploration of novel approaches, such as neurosymbolic AI (Garcez and Lamb, 2023), that have increased explainability and are operator-centred, to ensure these provide the intended decision support. Guidelines need to be developed to detect concept drift in a variety of representative manufacturing scenarios, and to decide between model updating and decommissioning of models that are no longer useful. Model verification in realistically designed environments that capture the stochasticity of manufacturing scenarios, knowledge-based AI verification, and test case generation are further areas that are in need of attention. Open source algorithms and public datasets can accelerate research in this area.

6.3. Risks pertaining to failures in trustworthy AI practice

Much technical research needs to be undertaken to identify and remove biases in data, interpret algorithmic results, verify models before deploying them, and detect when an algorithm needs to be updated or decommissioned. However, implementation of these technical advances is dependent on organizational practices that not only allow but also encourage such scrutiny and corrections to take place. We raised several research questions that are in need of attention relating to the right balance between regulation and innovation, and the practical implementation of regulatory frameworks or standards. There is currently no clear consensus on the types of data collection and algorithmic surveillance that would constitute unfair outcomes. A further complicating factor for manufacturers whose supply chains span multiple jurisdictions is potentially differing AI standards. The outsourcing of parts of the AI lifecycle may yield further issues where responsibility of development and verification is dispersed and perhaps untraceable. Organizations need to develop robust mechanisms to ensure safety from adversarial attacks and to improve ethical practices. Wide-reaching, visible leadership commitment should be considered with practices appropriate to the organizational setting, such as ethical review boards, publicly available responsible AI reports, whistleblowing roles, and reward mechanisms to increase incentives for trustworthy AI practice.

It is also worth noting that while this paper presents the first comprehensive framing of trustworthy AI in manufacturing and highlights the risks associated with AI, it represents only a first step in this increasingly important area. Both the field of AI and AI adoption in manufacturing are rapidly evolving, as do the definitions of what constitutes trustworthy AI. As we have seen, terminology pertaining to

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

ethics, trustworthiness, and responsible AI is not yet agreed upon, although there is emergent consensus on the main high-level principles, such as fairness, accountability, safety, and responsible use that is beneficial to humans. Researchers working at the intersection of AI and manufacturing engineering need to not only familiarize themselves with these developments but also take an active role in shaping and interpreting them for manufacturing. More illustrative use cases and best practices need to be brought to light in order to guide research-informed practice.

Looking ahead, several avenues for future research emerge from this study. First, while this paper maps the conceptual and practical dimensions of trustworthy AI in manufacturing and supply chains, further empirical work is needed to examine how organizations actually interpret and implement these principles in real-world settings. In particular, in-depth case analyses could provide insight into organizational decision-making, capability development, and governance practices. Second, the mapping has revealed gaps between high-level frameworks and the day-to-day concerns of manufacturing firms; future work should focus on developing tailored, lightweight tools and guidance to bridge this divide. Third, our taxonomy and synthesis can serve as a foundation for designing operational metrics and assessment mechanisms to evaluate trustworthiness in AI systems; this would require collaborative efforts between researchers, standards bodies, and industry. Finally, as generative AI and foundation models become more prevalent in industrial contexts, their alignment with principles of trustworthiness remains underexplored. We therefore encourage future work to examine how emerging technologies challenge, reinforce, or reconfigure current thinking in this area.

Data availability statement. Data sharing is not applicable to this article as no new data were created or analysed in this study.

Author contribution. Conceptualization-Equal: G.B., A.T., S.R., G.M-A., J.S.; Conceptualization-Lead: A.B.; Formal analysis-Equal: G.M-A.; Formal analysis-Lead: A.B., G.B.; Formal analysis-Supporting: J.S.; Investigation-Equal: G.M-A.; Investigation-Lead: A.B., G.B.; Investigation-Supporting: A.T., S.R., J.S.; Methodology-Equal: G.M-A.; Methodology-Lead: A.B., G.B.; Resources-Equal: A.B., G.B., G.M-A.; Visualization-Lead: G.M-A.; Visualization-Supporting: A.B., G.B.; Writing – Original Draft-Equal: G.M-A.; Writing – Original Draft-Lead: A.B., G.B.; Writing – Original Draft-Supporting: A.T., S.R., J.S.; Writing – Review & Editing-Equal: G.M-A.; Writing – Review & Editing-Lead: A.B., G.B.

Funding statement. Alexandra Brintrup was supported by The Alan Turing Institute's Data Centric Engineering Programme under the Lloyd's Register Foundation grant GBQ-100004.

Competing interests. The authors declare none.

Ethical standard. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

References

Adadi A and Berrada M (2018) Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052.

Adler R, **Feth P and Schneider D** (2016) Safety engineering for autonomous vehicles. In *2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshop (DSN-W)*, pp. 200–205.

Alaa AM and van der Schaar M (2019) Demystifying black-box models with symbolic metamodels. In Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E and Garnett R (eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc, pp. 11304–11314

Ali S, **Abuhmed T**, **El-Sappagh S**, **Muhammad K**, **Alonso-Moral J**, **Confalonieri R**, **Guidotti R**, **Del Ser J**, **Diaz-Rodriguez N and Herrera F** (2023) Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*. https://doi.org/10.1016/j.inffus.2023.101805.

Ananny M and Crawford K (2018) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society 20*(3), 973–989. https://doi.org/10.1177/1461444816676645.

Antoniou G, **Papadakis E and Baryannis G** (2022) Mental health diagnosis: A case for explainable artificial intelligence. *International Journal on Artificial Intelligence Tools 31*(03), 2241003. https://doi.org/10.1142/S0218213022410032.

Arrieta AB, **Díaz-Rodríguez N**, **Del Ser J**, **Bennetot A**, **Tabik S**, **Barbado A**, **García S**, **Gil-López S**, **Molina D**, **Benjamins R**, *et al.* (2020) Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

Ashmore R, **Calinescu R and Paterson C** (2021) Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ACM Computing Surveys (CSUR) 54*(5), 1–39.

Balasubramaniam N, **Kauppinen M**, **Rannisto A**, **Hiekkanen K and Kujala S** (2023) *Transparency and Explainability of AI Systems: From Ethical Guidelines to Requirements*. Information and Software Technology. https://doi.org/10.1016/j.infsof.2023.107197

Baryannis G, **Dani S and Antoniou G** (2019) Predicting supply chain risks using machine learning: The trade-off between performance and interpretability. *Future Generation Computer Systems 101*, 993–1004.

Baryannis G, **Kritikos K and Plexousakis D** (2017) A specification-based QoS-aware design framework for service-based applications. *Service Oriented Computing and Applications 11*(3), 301–314. https://doi.org/10.1007/s11761-017-0210-4.

Baryannis G and Plexousakis D (2013) WSSL: A fluent calculus-based language for web service specifications. In Salinesi C, Norrie MC and Pastor Ó (eds.), *25th International Conference on Advanced Information Systems Engineering (CAiSE 2013)*. Berlin/Heidelberg: Springer, pp. 256–271. https://doi.org/10.1007/978-3-642-38709-8_17

Baryannis, G., & **Plexousakis, D.** (2014). Fluent calculus-based semantic web service composition and verification using WSSL. In A. Lomuscio, et al. (eds.), *9th International Workshop on Semantic Web Enabled Software Engineering (SWESE2013), Co-Located with ICSOC 2013*. Switzerland: Springer International Publishing, pp. 256–270. https://doi.org/10.1007/978-3-319-06859-6_23

Baryannis G, **Validi S**, **Dani S and Antoniou G** (2018) Supply chain risk management and artificial intelligence: State of the art and future research directions. *International Journal of Production Research 59*(7), 2179–2202. https://doi.org/10.1080/00207543.2018.1530476.

Besinger P, **Vejnoska D and Ansari F** (2024) Responsible ai (rai) in manufacturing: A qualitative framework [5th international conference on industry 4.0 and Smart manufacturing (ISM 2023)]. *Procedia Computer Science 232*, 813–822. https://doi.org/10.1016/j.procs.2024.01.081.

Bibel W (2013) *Automated Theorem Proving*. Springer Science & Business Media

Borg M, **Englund C**, **Wnuk K**, **Durán B**, **Levandowski C**, **Gao S**, **Tan Y**, **Kaijser H**, **Lönn H and Törnqvist J** (2018) Safely entering the deep: A review of verification and validation for machine learning and a challenge elicitation in the automotive industry. *CoRR, abs/1812.05389*, 1–29. https://doi.org/10.48550/arXiv.1812.05389

Bostrom A, **Demuth J**, **Wirz C**, **Cains M**, **Schumacher A**, **Madlambayan D**, **Bansal A**, **Bearth A**, **Chase R**, **Crosman K**, **Ebert-Uphoff I**, **Gagne I**, **Guikema S**, **Hoffman R**, **Johnson B**, **Kumler-Bonfanti C**, **Lee J**, **Lowe A**, **McGovern A and Williams J** (2024) Trust and trustworthy artificial intelligence: A research agenda for ai in the environmental sciences. *Risk Analysis 44*(6), pp.1498–1513. https://doi.org/10.1111/risa.14245.

Brintrup A, **Kosasih EE**, **MacCarthy BL and Demirel G** (2022) Digital supply chain surveillance: Concepts, challenges, and frameworks. In *The Digital Supply Chain*. Elsevier, pp. 379–396.

Brundage M, **Avin S**, **Wang J**, **Belfield H**, **Krueger G**, **Hadfield GK**, **Khlaaf H**, **Yang J**, **Toner H**, **Fong R**, **Maharaj T**, **Koh PW**, **Hooker S**, **Leung J**, **Trask A**, **Bluemke E**, **Lebensold J**, **O'Keefe C**, **Koren M**, *et al* (2020) Toward trustworthy AI development: Mechanisms for supporting verifiable claims.

Chiusi F (2020) *Automating Society Report 2020* (Tech. rep.). AlgorithmWatch.

Clark M,**Kearns K**, **Overholt J**, **Gross K**, **Barthelemy B and Reed C** (2014) *Air Force Research Laboratory Test and Evaluation, Verification and Validation of Autonomous Systems Challenge Exploration* (tech. rep.). Air Force Research Lab.

Clarke EM, **Grumberg O**, **Kroening D**, **Peled D and Veith H** (2018) *Model Checking*. MIT press

Cobbe J, **Lee MSA and Singh J** (2021) Reviewable automated decision-making: A framework for accountable algorithmic systems. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* pp. 598–609. https://doi.org/10.1145/3442188.3445921.

Cobbe J and Singh J (2021) Artificial intelligence as a service: Legal responsibilities, liabilities, and policy challenges. *Computer Law & Security Review 42*, 105573. https://doi.org/10.1016/j.clsr.2021.105573.

Cobbe J, **Veale M and Singh J** (2023) Understanding accountability in algorithmic supply chains. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1186–1197. https://doi.org/10.1145/3593013.3594073.

Corrêa N, **Galvão C**, **Santos J**, **Del Pino C**, **Pinto E**, **Barbosa C**, **Massmann D**, **Mambrini R**, **Galvão L**, **Terem E and de Oliveira N** (2023) Worldwide ai ethics: A review of 200 guidelines and recommendations for ai governance. *Patterns*. https://doi.org/10.1016/j.patter.2023.100857.

Davis J, **Biller S**, **St Pierre JA and Jahanmir S** (2022) *Towards Resilient Manufacturing Ecosystems through Artificial Intelligence - Symposium Report (tech. rep.)*. National Institute of Standards and Technology NIST AMS. http://doi.org/10.6028/NIST.AMS.100-47

DCMS (2022, January 12) AI activity in UK businesses. Retrieved March 29, 2023, from https://www.gov.uk/government/publications/ai-activity-in-uk-businesses.

De Cremer D and Stollberger J (2022) Are people analytics dehumanizing your employees? *Harvard Business Review 2022*.

De Laat PB (2018) Algorithmic decision-making based on machine learning from big data: Can transparency restore accountability? *Philosophy & Technology 31*(4), 525–541.

Deng WH, **Nagireddy M**, **Lee MSA**, **Singh J**, **Wu ZS**, **Holstein K and Zhu H** (2022) Exploring how machine learning practitioners (try to) use fairness toolkits. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 473–484.

Department for Science, Innovation and Technology (2023) A pro-innovation approach to AI regulation. Presented to Parliament by the Secretary of State for Science, Innovation and Technology by Command of His Majesty on 29 March 2023.Command Paper Number: 815, ISBN: 978-1-5286-4009-1.

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

Dhada M, **Jain AK**, **Herrera M**, **Perez Hernandez M and Parlikad AK** (2020) Secure and communications-efficient collaborative prognosis. *IET Collaborative Intelligent Manufacturing 2*(4), 164–173.

Díaz-Rodríguez N, **Del Ser J**, **Coeckelbergh M**, **López de Prado M**, **Herrera-Viedma E and Herrera F** (2023) Connecting the dots in trustworthy artificial intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion 99*, 101896. https://doi.org/10.1016/j.inffus.2023.101896.

Dignum V (2023) Responsible artificial intelligence—From principles to practice: A keynote at thewebconf 2022. *ACM SIGIR Forum 56*, 1–6.

Ding W, **Abdel-Basset M**, **Hawash H and Ali A** (2022) Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey. *Information Sciences*, *615*, pp. 238–292. https://doi.org/10.1016/j.ins.2022.10.013.

Du J and Ling CX (2010) Active learning with human-like noisy oracle. *2010 IEEE International Conference on Data Mining*. pp. 797–802.

Dulac-Arnold G, **Levine N**, **Mankowitz DJ**, **Li J**, **Paduraru C**, **Gowal S and Hester T** (2021) Challenges of real-world reinforcement learning: Definitions, benchmarks and analysis. *Machine Learning 110*(9), 2419–2468.

Dwork C (2006) Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium*, ICALP 2006, Venice, Italy, July 10–14, 2006, Proceedings, Part II 33, pp. 1–12.

El Mhamdi EM, **Guerraoui R and Rouault S** (2017) On the robustness of a neural network. *IEEE 36th Symposium on Reliable Distributed Systems (SRDS) 2017*, 84–93. http://doi.org/10.1109/SRDS.2017.21.

Elendu C, **Amaechi D**, **Elendu T**, **Jingwa K**, **Okoye O**, **John Okah M**, **Ladele J**, **Farah A and Alimi H** (2023) Ethical implications of ai and robotics in healthcare: A review. *Medicine (United States)*. https://doi.org/10.1097/MD.0000000000036671.

European Commission. (2021). *Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts* (tech. rep.). European Commission.

Farahani B and Monsefi AK (2023) Smart and collaborative industrial iot: A federated learning and data space approach. *Digital Communications and Networks*, *9*(2), 436–447.

Fathy Y, **Jaber M and Brintrup A** (2020) Learning with imbalanced data in smart manufacturing: A comparative analysis. *IEEE Access 9*, 2734–2757.

Felzmann H, **Villaronga E**, **Lutz C and Tamò-Larrieux A** (2019) Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data and Society*. https://doi.org/10.1177/2053951719860542.

Floridi L (2019) Establishing the rules for building trustworthy ai. *Nature Machine Intelligence 1*, 261–262.

Floridi L, **Cowls J**, **Beltrametti M**, **Chatila R**, **Chazerand P**, **Dignum V**, **Luetge C**, **Madelin R**, **Pagallo U**, **Rossi F**, **Schafer B**, **Valcke P and Vayena E** (2018) Ai4people—An ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and Machines*. https://doi.org/10.1007/s11023-018-9482-5.

Floridi L, **Cowls J**, **King T and Taddeo M** (2020) How to design ai for social good: Seven essential factors. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-020-00213-5.

Floridi L and Taddeo M (2016) What is data ethics? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083), p.20160360. https://doi.org/10.1098/rsta.2016.0360.

Garcez A d and Lamb LC (2023) Neurosymbolic AI: The 3rd wave. *Artificial Intelligence Review*, 1–20.

Gentry C (2009) Fully homomorphic encryption using ideal lattices. *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, 169–178.

Géron A (2019) *Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media

Giovanola B and Tiribelli S (2023) Beyond bias and discrimination: Redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. *AI and Society*. https://doi.org/10.1007/s00146-022-01455-6.

Glimm B, **Horrocks I**, **Motik B**, **Stoilos G and Wang Z** (2014) Hermit: An OWL 2 reasoner. *Journal of Automated Reasoning 53*, 245–269.

Gu T, **Dolan-Gavitt B and Garg S** (2017) Badnets: Identifying vulnerabilities in the machine learning model supply chain. *CoRR abs/1708.06733* 1–13. http://arxiv.org/abs/1708.06733

Hagendorff T (2020) The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines 30*(1), 99–120.

Hanley D and Hubbard S (2020) *Eyes Everywhere: Amazon's Surveillance Infrastructure and Revitalizing Worker Power*. Open Markets Institute

Hansson K, **Yella S**, **Dougherty M and Fleyeh H** (2016) Machine learning algorithms in heavy process manufacturing. *American Journal of Intelligent Systems 6*(1), 1–13.

Hermann E (2022) Leveraging artificial intelligence in marketing for social good—An ethical perspective. *Journal of Business Ethics*. https://doi.org/10.1007/s10551-021-04843-y.

Hern A (2023, March 29) *Elon Musk Joins Call for Pause in Creation of Giant AI 'Digital Minds'*. Retrieved March 29, 2023, from https://www.theguardian.com/technology/2023/mar/29/elon-musk-joins-call-for-pause-in-creation-of-giant-aidigital-minds

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics Guidelines for Trustworthy AI* (tech. rep.). European Commission.

Hosseini S and Ivanov D (2020) Bayesian networks for supply chain risk, resilience and ripple effect analysis: A literature review. *Expert Systems with Applications 161*, 113649.

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

Huynh TD, **Tsakalakis N**, **Helal A**, **Stalla-Bourdillon S and Moreau L** (2021) Addressing regulatory requirements on explanations for automated decisions with provenance—A case study. *Digital Government: Research and Practice 2*(2). https://doi.org/10.1145/3436897.

Jobin A, **Ienca M and Vayena E** (2019) The global landscape of AI ethics guidelines. *Nature Machine Intelligence 1*(9), 389–399.

Joyce D, **Kormilitzin A**, **Smith K and Cipriani A** (2023) Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *npj Digital Medicine*. https://doi.org/10.1038/s41746-023-00751-9.

Kadadi A, **Agrawal R**, **Nyamful C and Atiq R** (2014) Challenges of data integration and interoperability in big data. In *2014 IEEE International Conference on Big Data (Big Data)*, pp. 38–40. https://doi.org/10.1109/BigData.2014.7004486.

Kattnig M, **Angerschmid A**, **Reichel T and Kern R** (2024) Assessing trustworthy AI: Technical and legal perspectives of fairness in AI. *Computer Law and Security Review*. https://doi.org/10.1016/j.clsr.2024.106053.

Kaur D, **Uslu S**, **Rittichier KJ and Durresi A** (2022) Trustworthy artificial intelligence: A review. *ACM Computing Surveys (CSUR) 55*(2), 1–38.

Kazim E and Koshiyama A (2021) A high-level overview of AI ethics. *Patterns*. https://doi.org/10.1016/j.patter.2021.100314.

Khan A, **Akbar M**, **Fahmideh M**, **Liang P**, **Waseem M**, **Ahmad A**, **Niazi M and Abrahamsson P** (2023) AI ethics: An empirical study on the views of practitioners and lawmakers. *IEEE Transactions on Computational Social Systems*. https://doi.org/10.1109/TCSS.2023.3251729.

Kordzadeh N and Ghasemaghaei M (2022) Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*. https://doi.org/10.1080/0960085X.2021.1927212.

Kosasih EE and Brintrup A (2022) Reinforcement learning provides a flexible approach for realistic supply chain safety stock optimisation. *IFAC-PapersOnLine 55*(10), 1539–1544.

Krauß J, **Pacheco BM**, **Zang HM and Schmitt RH** (2020) Automated machine learning for predictive quality in production. *Procedia CIRP 93*, 443–448.

Kroll JA, **Huey J**, **Barocas S**, **Felten EW**, **Reidenberg JR**, **Robinson DG and Yu H** (2017) Accountable algorithms. *University of Pennsylvania Law Review 165*.

Kumar RSS, **Nyström M**, **Lambert J**, **Marshall A**, **Goertzel M**, **Comissoneru A**, **Swann M and Xia S** (2020) Adversarial machine learning-industry perspectives. *IEEE Security and Privacy Workshops (SPW) 2020*, 69–75.

Laux J, **Wachter S and Mittelstadt B** (2024) Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation and Governance*. https://doi.org/10.1111/rego.12512.

Law R, **Ye H and Lei S** (2025) Ethical artificial intelligence (AI): Principles and practices. *International Journal of Contemporary Hospitality Management*. https://doi.org/10.1108/IJCHM-04-2024-0482.

Legg S and Hutter M (2007) Universal intelligence: A definition of machine intelligence. *Minds and machines*, *17*(4), pp. 391–444.

Lee MSA and Singh J (2021) The landscape and gaps in open source fairness toolkits. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3411764.3445261.

Leidner JL and Plachouras V (2017) Ethical by design: Ethics best practices for natural language processing. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 30–40.

Lewicki K, **Lee MSA**, **Cobbe J and Singh J** (2023) Out of context: Investigating the bias and fairness concerns of "artificial intelligence as a service". *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–17. https://doi.org/10.1145/3544548.3581463.

Lewis P and Marsh S (2022) What is it like to trust a rock? A functionalist perspective on trust and trustworthiness in artificial intelligence. *Cognitive Systems Research*. https://doi.org/10.1016/j.cogsys.2021.11.001.

Li B, **Qi P**, **Liu B**, **Di S**, **Liu J**, **Pei J**, **Yi J and Zhou B** (2023) Trustworthy AI: From principles to practices. *ACM Computing Surveys*. https://doi.org/10.1145/3555803.

Li X, **Xiong H**, **Li X**, **Wu X**, **Zhang X**, **Liu J**, **Bian J and Dou D** (2022) Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems*. https://doi.org/10.1007/s10115-022-01756-8.

Maher M, **Khan I and Prikshat V** (2023) Monetisation of digital health data through a GDPR-compliant and blockchain enabled digital health data marketplace: A proposal to enhance patient's engagement with health data repositories. *International Journal of Information Management Data Insights 3*(1), 100159.

Mak S, **Xu L**, **Pearce T**, **Ostroumov M and Brintrup A** (2021) *Coalitional Bargaining via Reinforcement Learning: An Application to Collaborative Vehicle Routing*. NeurIPS

Marcinkevičs R and Vogt J (2023) Interpretable and explainable machine learning: A methods-centric overview with concrete examples. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. https://doi.org/10.1002/widm.1493.

Markus A, **Kors J and Rijnbeek P** (2021) The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*. https://doi.org/10.1016/j.jbi.2020.103655.

Martin K (2019) Ethical implications and accountability of algorithms. *Journal of Business Ethics*. https://doi.org/10.1007/s10551-018-3921-3.

McKinsey and Company (2020) *Monetizing Car Data New Service Business Opportunities to Create New Customer Benefits.*

Mennella C, **Maniscalco U**, **De Pietro G and Esposito M** (2024) Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. *Heliyon*. https://doi.org/10.1016/j.heliyon.2024.e26297.

Mentzas G, **Fikardos M**, **Lepenioti K and Apostolou D** (2024) Exploring the landscape of trustworthy artificial intelligence: Status and challenges. *Intelligent Decision Technologies*. https://doi.org/10.3233/IDT-240366.

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

Mittelstadt BD, **Allo P**, **Taddeo M**, **Wachter S and Floridi L** (2016) The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679.

Morley J, **Kinsey L**, **Elhalal A**, **Garcia F**, **Ziosi M and Floridi L** (2023) Operationalising AI ethics: Barriers, enablers and next steps. *AI and Society*. https://doi.org/10.1007/s00146-021-01308-8.

Morley J, **Machado CC**, **Burr C**, **Cowls J**, **Joshi I**, **Taddeo M and Floridi L** (2020) The ethics of ai in health care: A mapping review. *Social Science & Medicine 260*, 113172.

Naz F, **Agrawal R**, **Kumar A**, **Gunasekaran A**, **Majumdar A and Luthra S** (2022) Reviewing the applications of artificial intelligence in sustainable supply chains: Exploring research propositions for future directions. *Business Strategy and the Environment 31*(5), 2400–2423.

Newman J (2023) *A Taxonomy of Trustworthiness for Artificial Intelligence* (tech. rep.). UC Berkeley Center for Long-Term Cybersecurity.

Nguyen A, **Ngo H**, **Hong Y**, **Dang B and Nguyen B-P** (2023) Ethical principles for artificial intelligence in education. *Education and Information Technologies*. https://doi.org/10.1007/s10639-022-11316-w.

Ntoutsi E, **Fafalios P**, **Gadiraju U**, **Iosifidis V**, **Nejdl W**, **Vidal M-E**, **Ruggieri S**, **Turini F**, **Papadopoulos S**, **Krasanakis E**, **Kompatsiaris I**, **Kinder-Kurlanda K**, **Wagner C**, **Karimi F**, **Fernandez M**, **Alani H**, **Berendt B**, **Kruegel T**, **Heinze C**, **Broelemann K**, **Kasneci G**, **Tiropanis T and Staab S** (2020) Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. https://doi.org/10.1002/widm.1356.

Office of Science and Technology Policy (2022) *Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People* (tech. rep.). The White House.

Oliff H, **Liu Y**, **Kumar M**, **Williams M and Ryan M** (2020) Reinforcement learning for facilitating human-robot-interaction in manufacturing. *Journal of Manufacturing Systems 56*, 326–340.

Owen R, **Macnaghten P and Stilgoe J** (2012) Responsible research and innovation: From science in society to science for society, with society. *Science and Public Policy 39*(6), 751–760. https://doi.org/10.1093/scipol/scs093.

Ozmen Garibay O, **Winslow B**, **Andolina S**, **Antona M**, **Bodenschatz A**, **Coursaris C**, **Falco G**, **Fiore S**, **Garibay I**, **Grieman K**, **Havens J**, **Jirotka M**, **Kacorri H**, **Karwowski W**, **Kider J**, **Konstan J**, **Koon S**, **Lopez-Gonzalez M**, **Maifeld-Carucci I**, **McGregor S**, **Salvendy G**, **Shneiderman B**, **Stephanidis C**, **Strobel C**, **ten Holter C and Xu W** (2023) Six human-centered artificial intelligence grand challenges. *International Journal of Human-Computer Interaction*. https://doi.org/10.1080/10447318.2022.2153320.

Paleyes A, **Urma R-G and Lawrence ND** (2022) Challenges in deploying machine learning: A survey of case studies. *ACM Computing Surveys 55*(6), 1–29.

Paraman P and Anamalah S (2023) Ethical artificial intelligence framework for a good AI society: Principles, opportunities and perils. *AI and Society*. https://doi.org/10.1007/s00146-022-01458-3.

Pasquier T, **Singh J**, **Powles J**, **Eyers D**, **Seltzer M and Bacon J** (2018) Data provenance to audit compliance with privacy policy in the internet of things. *Personal Ubiquitous Comput. 22*(2), 333–344. https://doi.org/10.1007/s00779-017-1067-4.

Pauwels P, **Zhang S and Lee Y-C** (2017) Semantic web technologies in AEC industry: A literature overview. *Automation in Construction 73*, 145–165. http://doi.org/10.1016/j.autcon.2016.10.003.

Petersen K, **Feldt R**, **Mujtaba S and Mattsson M** (2008) Systematic mapping studies in software engineering. In *EASE'08: Proceedings of the 12th international conference on Evaluation and Assessment in Software Engineering*. pp. 68–77.

Petersen K, **Vakkalanka S and Kuzniarz L** (2015) Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology 64*, 1–18. https://doi.org/10.1016/j.infsof.2015.03.007.

Peyre J, **Sivic J**, **Laptev I and Schmid C** (2017) Weakly-supervised learning of visual relations. In *Proceedings of the IEEE International Conference on Computer Vision*. pp. 5179–5188.

Pranckuté R (2021) Web of science (WOS) and Scopus: The titans of bibliographic information in today's academic world. *Publications 9*(1). https://doi.org/10.3390/publications9010012.

Radanliev P, **Santos O**, **Brandon-Jones A and Joinson A** (2024) Ethics and responsible AI deployment. *Frontiers in Artificial Intelligence*. https://doi.org/10.3389/frai.2024.1377011.

Radclyffe C, **Ribeiro M and Wortham R** (2023) The assessment list for trustworthy artificial intelligence: A review and recommendations. *Frontiers in Artificial Intelligence*. https://doi.org/10.3389/frai.2023.1020592.

Raji ID, **Smart A**, **White RN**, **Mitchell M**, **Gebru T**, **Hutchinson B**, **Smith-Loud J**, **Theron D and Barnes P** (2020) Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. pp. 33–44. http://doi.org/10.1145/3351095.3372873.

Ryan M and Stahl B (2021) Artificial intelligence ethics guidelines for developers and users: Clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society*. https://doi.org/10.1108/JICES-12-2019-0138.

Salay R, **Queiroz R and Czarnecki K** (2017) An analysis of ISO 26262: Using machine learning safely in automotive software. *CoRR*, *abs/1709.02435*, 1–6. http://arxiv.org/abs/1709.02435.

Sambasivan N and Holbrook J (2018) Toward responsible AI for the next billion users. *Interactions 26*(1), 68–71.

Seshia SA, **Sadigh D and Sastry SS** (2022) Toward verified artificial intelligence. *Communications of the ACM 65*(7), 46–55.

Shneiderman B (2020) Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction 36*(6), 495–504.

Shneiderman B (2021) Responsible AI: Bridging from ethics to practice. *Communications of the ACM 64*(8), 32–35.

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

Shokri R, **Stronati M**, **Song C and Shmatikov V** (2017) Membership inference attacks against machine learning models. *IEEE Symposium on Security and Privacy (SP) 2017*, 3–18.

Singh J, **Cobbe J and Norval C** (2019) Decision provenance: Harnessing data flow for accountable systems. *IEEE Access 7*, 6562–6574. https://doi.org/10.1109/ACCESS.2018.2887201.

Sirin E, **Parsia B**, **Grau BC**, **Kalyanpur A and Katz Y** (2007) Pellet: A practical OWL-DL reasoner. *Journal of Web Semantics 5*(2), 51–53.

Smuha NA (2019) The EU approach to ethics guidelines for trustworthy artificial intelligence. *Computer Law Review International 20*(4), 97–106.

Stahl B (2023) Embedding responsibility in intelligent systems: From AI ethics to responsible AI ecosystems. *Scientific Reports*. https://doi.org/10.1038/s41598-023-34622-w.

Starke C, **Baleis J**, **Keller B and Marcinkowski F** (2022) Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data and Society*. https://doi.org/10.1177/20539517221115189.

Strubell E, **Ganesh A and McCallum A** (2020) Energy and policy considerations for modern deep learning research. *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7–12, 2020.* pp. 13693–13696. https://ojs.aaai.org/index.php/AAAI/article/view/7123

Sun Y, **Song Q**, **Gui X**, **Ma F and Wang T** (2023) *AutoML in the Wild: Obstacles, Workarounds, and Expectations.* pp. 1–15. https://doi.org/10.1145/3544548.3581082

Sunmola F and Baryannis G (2024) Artificial intelligence opportunities for resilient supply chains [18th IFAC symposium on information control problems in manufacturing INCOM 2024]. *IFAC-PapersOnLine 58*(19), 813–818. https://doi.org/10.1016/j.ifacol.2024.09.195.

Suresh, H., & **Guttag, J.** (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, pp. 1–9.

Tang L, **Li J and Fantus S** (2023) Medical artificial intelligence ethics: A systematic review of empirical studies. *DIGITAL HEALTH*. https://doi.org/10.1177/20552076231186064.

Taylor BJ (2006) *Methods and Procedures for the Verification and Validation of Artificial Neural Networks*. Springer Science & Business Media

Terziyan V, **Golovianko M and Gryshko S** (2018) Industry 4.0 intelligence under attack: From cognitive hack to data poisoning. *Cyber defence in Industry 4*, 110–125.

Thiebes S, **Lins S and Sunyaev A** (2021) Trustworthy artificial intelligence. *Electronic Markets 31*, 447–464.

Trocin C, **Mikalef P**, **Papamitsiou Z and Conboy K** (2021) Responsible ai for digital health: A synthesis and a research agenda. *Information Systems Frontiers*, 1–19.

Trustworthy and Responsible AI Resource Center (2023) *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (tech. rep.). National Institute of Science and Technology.

Ueda D, **Kakinuma T**, **Fujita S**, **Kamagata K**, **Fushimi Y**, **Ito R**, **Matsui Y**, **Nozaki T**, **Nakaura T**, **Fujima N**, **Tatsugami F**, **Yanagawa M**, **Hirata K**, **Yamada A**, **Tsuboyama T**, **Kawamura M**, **Fujioka T and Naganawa S** (2024) Fairness of artificial intelligence in healthcare: Review and recommendations. *Japanese Journal of Radiology*. https://doi.org/10.1007/s11604-023-01474-3.

UK (2018, June) Government Response to House of Lords Artificial Intelligence Select Committee's Report on AI in the UK: Ready, Willing and Able? https://www.parliament.uk/globalassets/documents/lords-committees/Artificial-Intelligence/AIGovernment-Response2.pdf

van Eck NJ and Waltman L (2010) Software survey: VOS viewer, a computer program for bibliometric mapping. *Scientometrics 84*(2), 523–538.

van Giffen B, **Herhausen D and Fahse T** (2022) Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research 144*, 93–106.

Vyhmeister E and Castane GG (2025) Tai-prm: Trustworthy AI—Project risk management framework towards industry 5.0. *AI and Ethics 5*(2), 819–839. https://doi.org/10.1007/s43681-023-00417-y.

W3C Provenance Incubator Group (2010) *What Is Provenance.*

Wang Z and Hong T (2020) Reinforcement learning for building controls: The opportunities and challenges. *Applied Energy 269*, 115036.

Wang C, **Liu S**, **Yang H**, **Guo J**, **Wu Y and Liu J** (2023) Ethical considerations of using chatgpt in health care. *Journal of Medical Internet Research*. https://doi.org/10.2196/48009.

Wang Y, **Xiong M and Olya H** (2020) Toward an understanding of responsible artificial intelligence practices. In *Proceedings of the 53rd Hawaii International Conference on System Sciences.* pp. 4962–4971.

Wexler J, **Pushkarna M**, **Bolukbasi T**, **Wattenberg M**, **Viégas F and Wilson J** (2019) The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics 26*(1), 56–65.

Williams R, **Cloete R**, **Cobbe J**, **Cottrill C**, **Edwards P**, **Markovic M**, **Naja I**, **Ryan F**, **Singh J and Pang W** (2022) From transparency to accountability of intelligent systems: Moving beyond aspirations. *Data & Policy 4*, e7.

Yampolskiy M, **Graves L**, **Gatlin J**, **Skjellum A and Yung M** (2021) What did you add to my additive manufacturing data?: Steganographic attacks on 3d printing files. In *Proceedings of the 24th International Symposium on Research in Attacks, Intrusions and Defenses*. pp. 266–281.

https://doi.org/10.1017/dce.2025.10032 Published online by Cambridge University Press

Yong BX and Brintrup A (2020) Multi agent system for machine learning under uncertainty in cyber physical manufacturing system. *Service Oriented, Holonic and Multi-agent Manufacturing Systems for Industry of the Future: Proceedings of SOHOMA 2019 9*, 244–257.

Yousefi N, **Tsianikas S and Coit DW** (2020) Reinforcement learning for dynamic condition-based maintenance of a system with individually repairable components. *Quality Engineering 32*(3), 388–408.

Zhang J and Zhang Z-M (2023) Ethics and governance of trustworthy medical artificial intelligence. *BMC Medical Informatics and Decision Making*. https://doi.org/10.1186/s12911-023-02103-9.

Zheng G, **Kong L and Brintrup A** (2023) Federated machine learning for privacy preserving, collective supply chain risk prediction. *International Journal of Production Research*, 1–18.

Cite this article: Brintrup A, Baryannis G, Tiwari A, Ratchev S, Martinez-Arellano G and Singh J (2025). Trustworthy, responsible and ethical artificial intelligence in manufacturing and supply chains: synthesis and emerging research questions. *Data-Centric Engineering*, 6, e53. doi:10.1017/dce.2025.10032