

TRUSTWORTHY AI FRAMEWORK

A COMPREHENSIVE REVIEW OF AI STANDARDS POLICIES AND
A PRACTICAL GUIDELINE TO THEIR APPLICATION IN MANUFACTURING



Acknowledgements

Dr. Yazan Qarout,

Senior Research Engineer, Digital Engineering, The Manufacturing Technology Centre Ltd

Mohammed Begg,

Research Engineer, Digital Engineering, The Manufacturing Technology Centre Ltd

Liam Fearon,

Graduate Research Engineer, Digital Engineering, The Manufacturing Technology Centre Ltd

David Russell,

Graduate Research Engineer, Digital Engineering, The Manufacturing Technology Centre Ltd

Dr. Nikita Pietrow,

Advanced Research Engineer, Digital Engineering, The Manufacturing Technology Centre Ltd

Dr. Mostafizur Rahman,

Chief Technologist – Artificial Intelligence, Digital Engineering, The Manufacturing Technology Centre Ltd

Dr. Stuart McLeod,

Technology Manager, Digital Engineering, The Manufacturing Technology Centre Ltd

Dr. Nandini Chakravorti,

Associate Director, Digital Engineering, The Manufacturing Technology Centre Ltd

Tom Winter,

Technology Manager, Digital Engineering, The Manufacturing Technology Centre Ltd

James Fortune,

Rolls-Royce

Thomas Ota,

AWE

INDUSTRIAL STEERING GROUP MEMBERS:

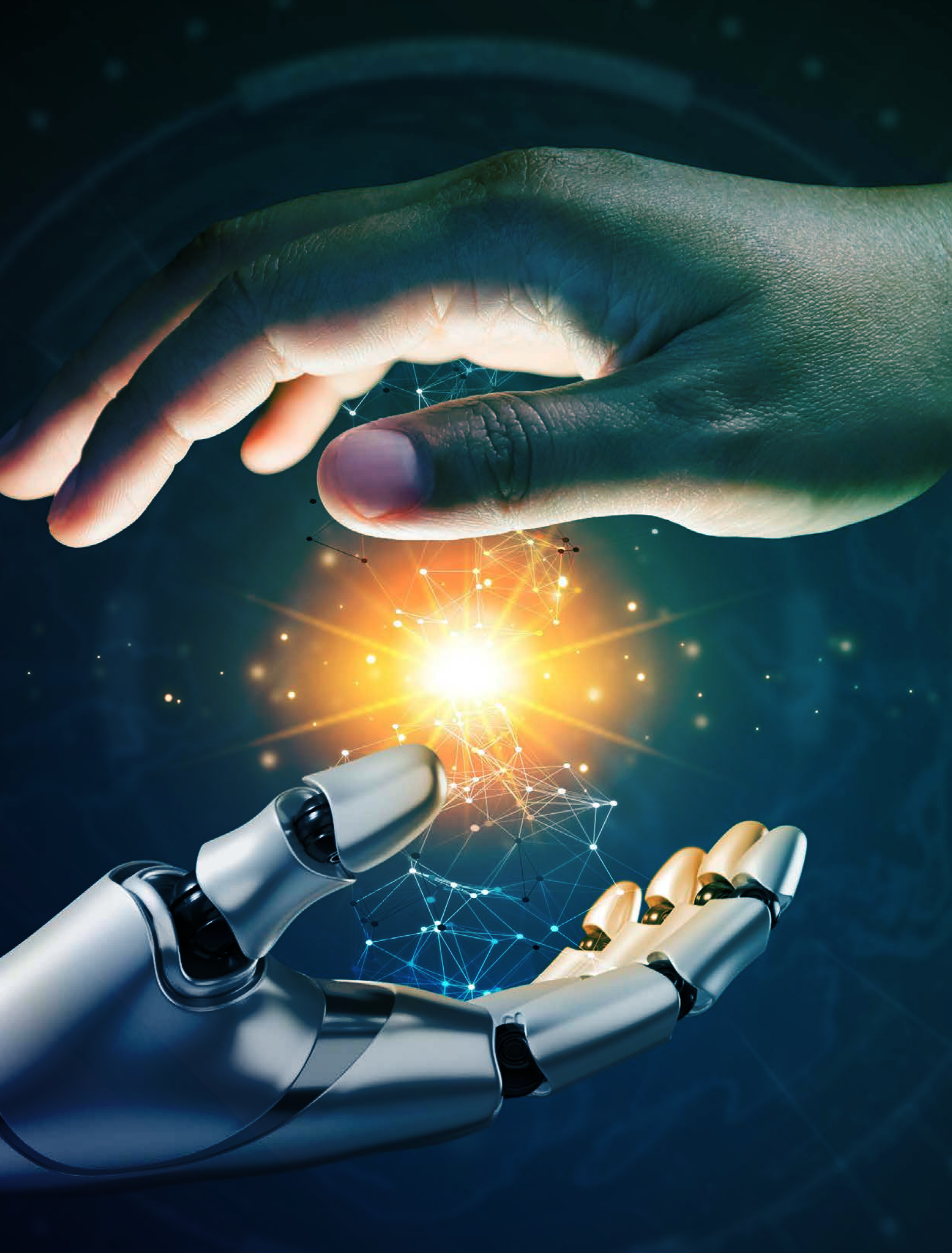


Members



Contents

| | |
|---|----|
| Acknowledgements | 5 |
| Executive Summary | 7 |
| Introduction | 8 |
| Trustworthy AI | 11 |
| Trustworthy AI in Manufacturing | 12 |
| AI Assurance Techniques | 15 |
| Current State of Trustworthy AI | 17 |
| Framework for Implementing Trustworthy AI | 19 |
| Conclusion | 20 |
| References | 22 |
| Appendix | 24 |



List of Figures

| | |
|---|----|
| Figure 1: General Challenges of AI Industrialisation. | 8 |
| Figure 2: The AI Development Lifecycle. | 13 |
| Figure 3: Trustworthiness principles | 13 |
| Figure 4: Diagram illustrating the pieces of reviewed literature that most comprehensively cover each trustworthy AI principle across the AI lifecycle | 14 |
| Figure 5: Diagram illustrating the pieces of reviewed literature that most comprehensively cover how to complete a particular AI assurance technique across the AI lifecycle. | 16 |
| Figure 6. An outline of the gaps in the AI standard literature and example of the developed trustworthy AI framework. | 21 |

List of Tables

| | |
|---|----|
| Table 1: Heatmap displaying the of coverage AI assurance techniques across the AI lifecycle in the literature of AI standards and frameworks. | 18 |
| Table 2: Heatmap displaying the of coverage trustworthy AI pillars across the AI lifecycle in the literature of AI standards and frameworks. | 18 |



Executive Summary

The progress of Artificial Intelligence (AI) in recent years has transformed the public perception of AI from being science fiction and researched by top universities to a tool for present day applications. While AI and Machine Learning technology has rapidly developed, the implementation and adoption of these technologies in the manufacturing sector has been slow. One of the reasons why this has occurred has been the lack of Trust in AI systems and clear understanding of standards and policies. Trustworthy AI looks to establish this trust through a series of principles and practices that provide confidence that an AI system is secure, fair, transparent, accountable, and robust. AI assurance techniques mean to provide a framework for manufacturing stakeholders to assess trustworthiness in the development and deployment of their AI systems.

Furthermore, while there are numerous AI standards and policies available, these policies are often difficult to extract relevant guidance making them impractical for AI developers. This information inundation makes it challenging for manufacturing stakeholders to discern which standards and policies are most relevant to their operations and which ones they should adhere to. As a result,

there is a pressing need for more comprehensive frameworks that offer clear, actionable recommendations for integrating AI technologies into manufacturing processes while ensuring trustworthiness and compliance with industry standards. Without such guidance, the potential benefits of AI in manufacturing may remain unrealised, hampering its widespread adoption and optimisation.

This paper provides a comprehensive view of current AI standards and frameworks and the current usage of these assurance techniques in generating trust in AI systems. The overall findings from the paper suggest that while current standards and frameworks demonstrate some of the assurance techniques, there is still yet to exist a holistic approach to these assurance techniques in AI systems. The recommendations for what are next in this area is to create an approach using the principles and practices and ensuring that this approach can be applied across a wider manufacturing sector.

The detailed framework is available as supplementary material upon request. To obtain your copy of the Trustworthy AI framework, please send your request to aigovernance@the-mtc.org.

Introduction

Artificial Intelligence (AI) has advanced rapidly in recent years as different developments have improved its capability and function, such as new and improved learning methods combined with the increased availability of large datasets for training and self-learning. The emergence of ChatGPT and other publicly available generative AI systems has made AI a very popular topic of discussion in the public sphere and garnered increased interest in many industry sectors as organisations look to harness the benefits of AI. However, AI has been employed in industry for some time in various forms with many identifying the potential advantages for different sectors and applications early on.

Manufacturing has been seen as an industry prime for AI implementation, with potential for significant value to be added across the whole value chain in a variety of ways. While AI is not a new concept, with much research and

work invested into this area, the wider adoption of AI within manufacturing sector is still in relatively early stages. In a survey conducted on UK businesses ^[1], just over 15% of the manufacturing business respondents indicated they have currently adopted AI, with roughly a further 13% intending to use it. A PwC global survey found just 9% of the 1,155 manufacturing executives surveyed had implemented AI into their processes to improve operational decision making ^[2]. These numbers are likely higher as of 2024, however the information highlights the fact that the manufacturing sector is not a leading AI adopter despite the identified benefits.

Although there are many opportunities and potential benefits to AI deployment for manufacturing, the industry remains slow to adopt the technology. Figure 1 shows the general challenges to the industrialisation of AI technologies from feedback of various vendors and end users across different sectors in manufacturing.

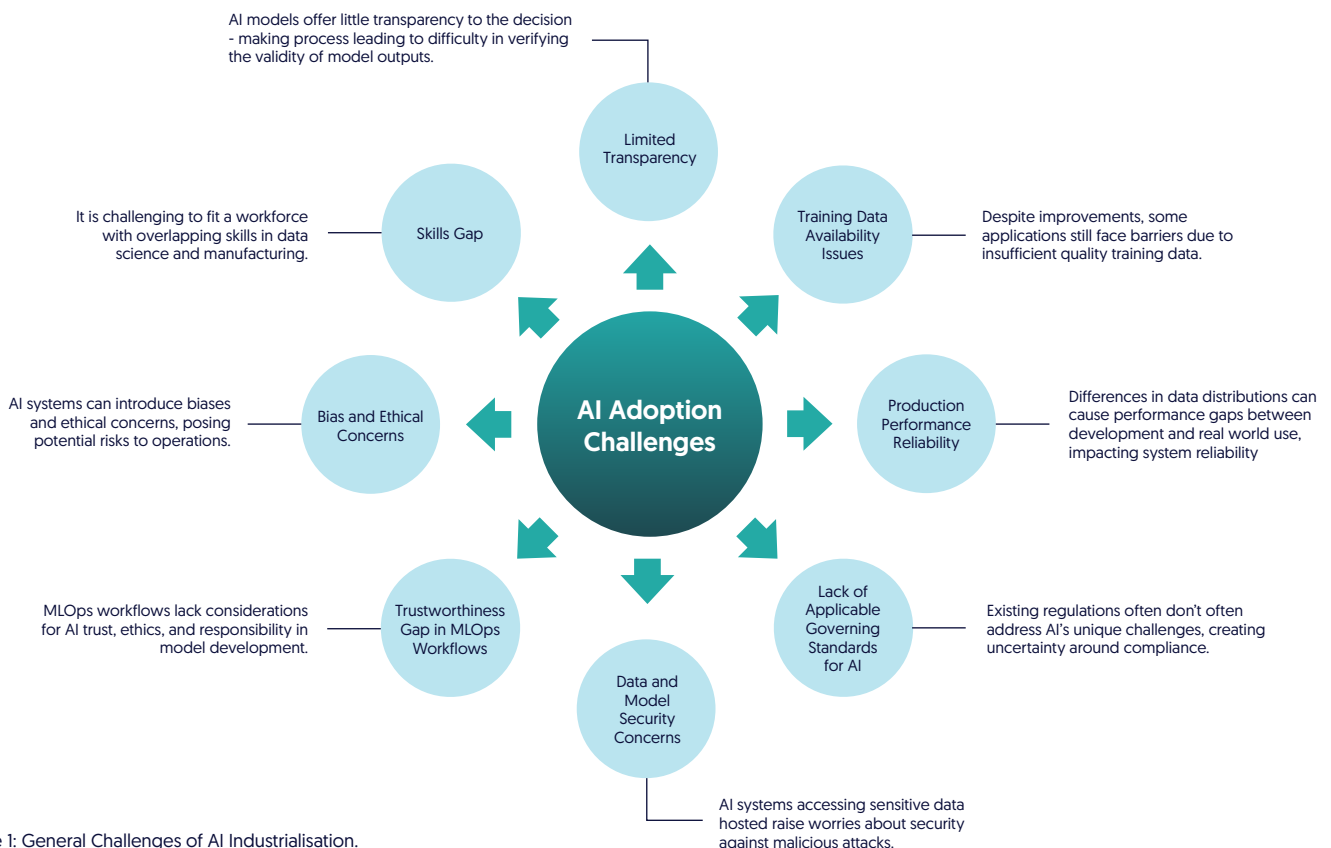


Figure 1: General Challenges of AI Industrialisation.

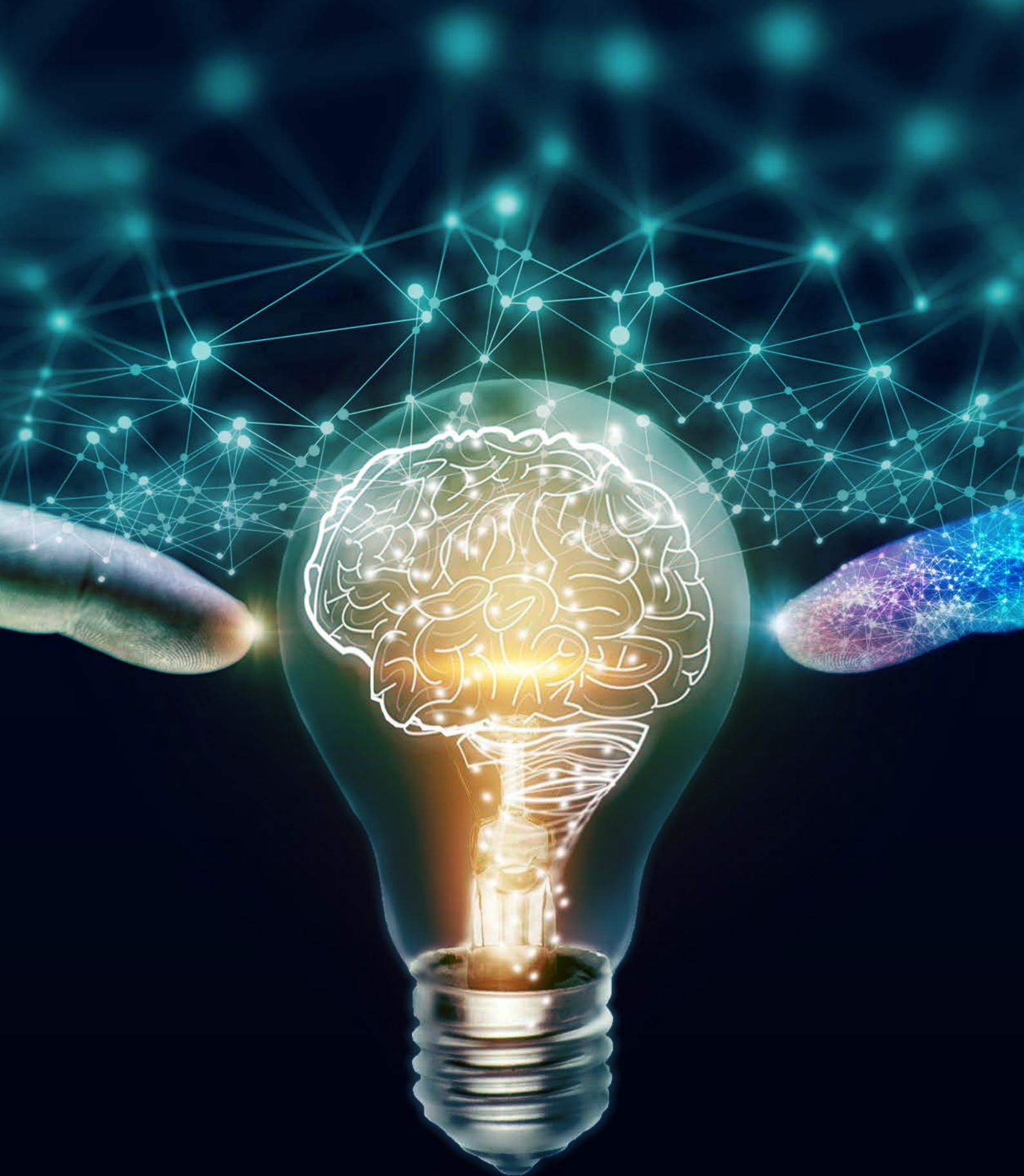
As seen in Figure 1 the concerns of the manufacturing sector stretch across a wide range of factors; from performance to ethical concerns, to a lack of transparency in decision making processes. Some of these challenges can be overcome by following Trustworthy AI principles which look at establishing guidance on development and deployment of AI models. There have been large amounts of research on this topic, including published papers and governmental guidance due to the rapid and continuous development of AI technology. However, it remains a poorly understood topic for the manufacturing industry with a lack of practical implementation of these principles through official frameworks.

Numerous national and international initiatives are aimed at addressing the topic of trustworthy AI, including the AI summit at Bletchley Park ^[3] and the EU AI act ^[4]. This has

encouraged an increased work effort on the standardisation of AI development and deployment, helping society make big strides in understanding AI best practices. Yet, with the large number of standards released in recent years, it is making it difficult for industry to navigate the plethora of frameworks available. This paper provides a structured guide to identifying the correct directives for the different themes within trustworthy AI.

This paper will further address the problem of implementing trustworthy AI for critical manufacturing application with an exploration of frameworks, standards, and reports covering the topic of trustworthy AI.





Trustworthy AI

Before delving into trustworthy AI, it's essential to establish a clear definition of AI itself. The MTC defines AI as leveraging data, algorithms, and machines to simulate intelligence, enabling them to perform tasks that require cognition such as decision making and problem solving. The noun Intelligence refers to the cognitive ability to perceive knowledge or information which is then applied to carry out functional tasks.

Implementation of AI can range from being used in computer vision systems to identify for quality assurance in manufacturing lines, to using multiple AI models in which predictive maintenance combined with a large language model can be used to inform manufacturers about problems and possible solutions in more human-centric way. In addition, AI's impact on manufacturing extends to a wide array of scenarios, such as revolutionising inventory management through real-time demand forecasting or facilitating personalised product customisation through adaptive production processes.

By leveraging AI in these diverse capacities, manufacturers can achieve unprecedented levels of agility, responsiveness, and customer satisfaction, ultimately driving competitive advantage in the rapidly evolving marketplace. Furthermore, AI finds application in various facets of manufacturing, spanning from optimising supply chain management through predictive analytics to enhancing robotic automation for streamlined assembly processes. By harnessing AI technologies in these diverse contexts, manufacturers can unlock efficiencies, improve productivity, and drive innovation across their operations.

WHAT IS TRUSTWORTHY AI?

Trustworthy AI refers to the development and deployment of AI systems that are secure, robust, transparent, and ethical, and explainable. The term often encompasses a set of principles and techniques aimed at ensuring that AI technology is developed, deployed, and operated in a manner that inspires confidence by the user, developer, and all associated stakeholders.

Trustworthy AI gained prominence with the rapid advancement of AI technology. Traction for trustworthiness began in the 2010s when AI adoption became a topic of importance across industries. Today, AI trust is a crucial topic for safeguarding against any harm that can come from the misuse of this technology and ensuring that an AI model is safe to operate. This lack of trust in AI acts as a bottleneck that is slowing adoption in many industries including manufacturing. Developing trustworthy AI models is an involved process requiring a variety of specialised roles throughout the entirety of the AI development lifecycle, as depicted in Figure 2 and explained in detail in the whitepaper 'Utilising Externally Hosted AI/Generative AI Services in Manufacturing' ^[5].

In line with the guidance provided by the UK government's Centre for Data Ethics and Innovation (CDEI) and the Department for Science, Innovation and Technology (DSIT) ^[6]. Trustworthy AI revolves around 5 core principles illustrated in Figure 3 and should adhere to important AI assurance techniques throughout the AI development lifecycle. The following sections discuss the respective principles and assurance techniques in more detail, while highlighting the existing standards and frameworks that adequately address them as of November 2023.

Trustworthy AI in Manufacturing

Trustworthy AI practices and principles are particularly critical for manufacturing applications, where incorrect integration of AI in production systems can be detrimental to productivity and harmful if used in an inappropriate manner. Below are some examples of AI applications in manufacturing and the importance trustworthy AI holds for them.

AI FOR QUALITY ASSURANCE

In the realm of quality control in manufacturing, AI is harnessed to detect defects, anomalies, and deviations in the production process. Trustworthy AI is of paramount importance in this context, as the reliability and accuracy of the AI system directly influence the quality of the final products. A trustworthy AI solution not only provides precise defect detection but also offers interpretability and explainability, allowing manufacturers to understand why a certain decision was made. This transparency is crucial, especially when product quality is synonymous with safety and regulatory compliance. Moreover, trustworthy AI in quality control ensures consistency and fairness in the inspection process, mitigating the risk of biased decisions that could compromise the reliability of the manufacturing process. This can occur when the AI model produces results that are systematically prejudiced to certain decisions, potentially due to inherent bias in the training dataset. As manufacturers increasingly wish to rely on AI to uphold quality standards, the trustworthy implementation of AI in quality control not only enhances operational efficiency but also fosters confidence in the reliability and fairness of the entire production chain. Companies collaborating across the supply chain can place trust in the consistent quality of AI-evaluated products received from their partners, thereby guaranteeing the quality of their final products.

PROCESS AND ROBOTIC AUTOMATION

Autonomous robots incorporate AI to automate manu-

facturing functions, traditionally for repetitive and monotonous tasks including picking and placing, welding, painting, assembly, and inspection that free up workers time to focus on more productive tasks. Integrating robots with machine learning allows them to self-learn and become better at tasks at hand or learn new tasks, analysing its own performance to improve over time. Machine vision can enhance capabilities further by enabling better mobility in complex environments while helping with safety when working around human colleagues. Robots in these situations are often referred to as collaborative robots (Cobots for short) as they work in support of their human counterpart and are specifically designed to augment the human worker. Extending this beyond individual tasks, whole processes can be automated, making use of automated robotics, machine learning, and business process automation to streamline the process from start to finish. Using AI enabled automation in production relies heavily on trustworthy AI practices. Ensuring reliability, safety, and transparency in their decision-making processes becomes pivotal, especially when integrating machine learning to enhance their capabilities. Trustworthy AI not only bolsters the efficiency of these systems but also instils confidence in their seamless collaboration with human counterparts, emphasising a safe and productive working environment.

LARGE LANGUAGE MODELS (LLMS)

In the manufacturing industry, conversational AI and Large Language Models (LLMs) play a pivotal role in simplifying access to unstructured data while offering an intuitive interface for information retrieval. These technologies streamline the interaction with complex datasets, enabling effortless querying of vast amounts of manufacturing-related information. Through chatbots or virtual assistants tailored for manufacturing contexts, conversational AI facilitates quick access to insights and operational details. Simultaneously, LLMs adeptly process textual manufacturing data, providing detailed responses

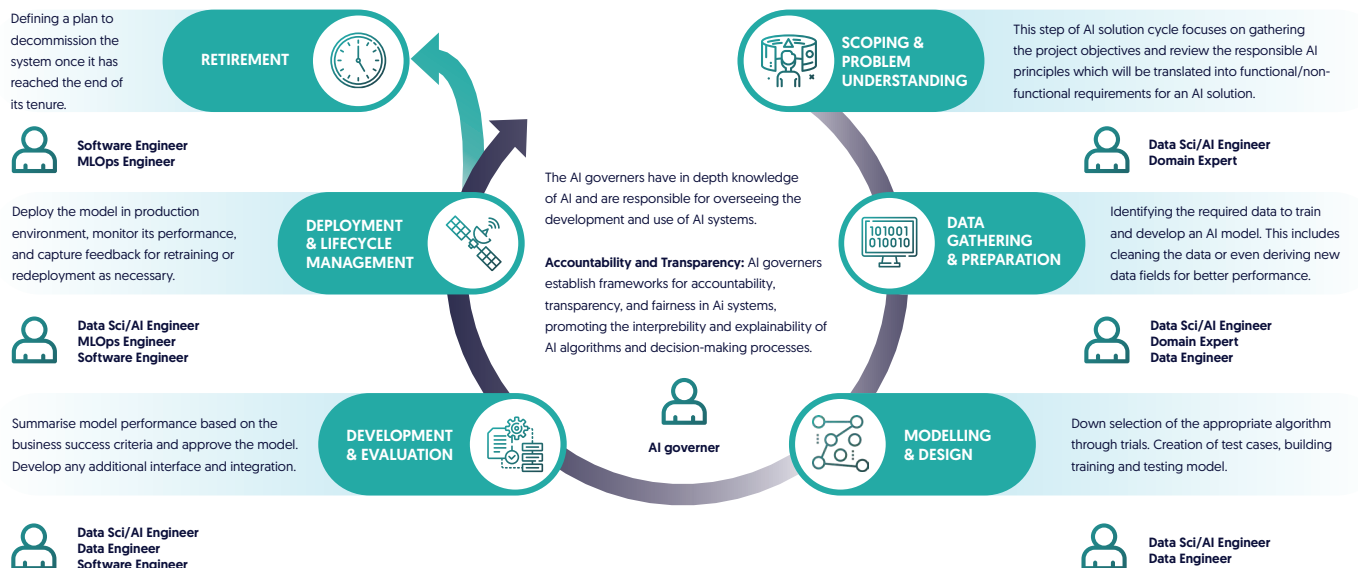


Figure 2: The AI Development Lifecycle^[5].

and improving comprehension of unstructured information specific to this sector. Hallucination and invention (the creation of false references where none exist for instance) are problems rooted deep in the heart of how LLMs are structured and trained. The boundary between reliable and unreliable output in an LLM is very complex and hard to discern. These models must be properly assessed against trustworthy AI principles to allow for adoption in critical applications.

TRUSTWORTHY AI PRINCIPLES

Trustworthy AI principles encompass a set of values that guide the design, development, deployment, and use of AI systems. These principles aim to ensure that AI technologies are developed and applied in a manner that respects governance guidelines, maintains data security, ensures human and asset safety, as well as considering appropriate model transparency.

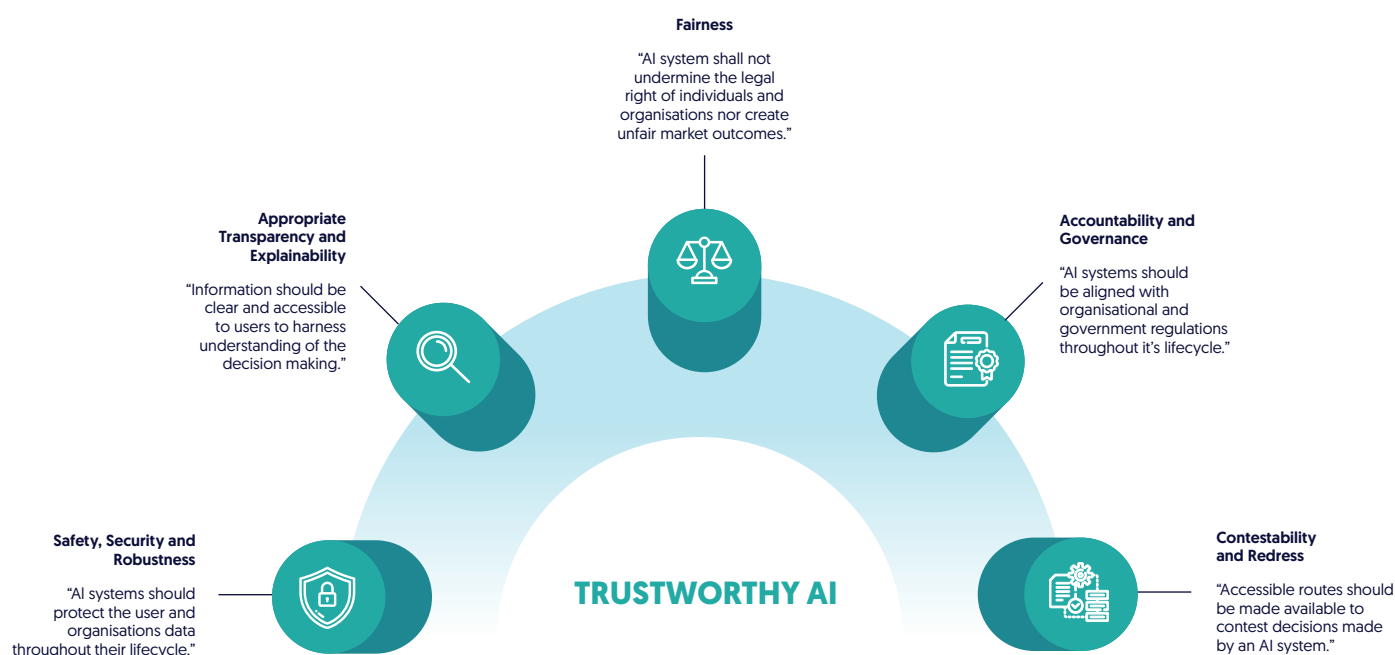


Figure 3: Trustworthy principles.

These principles were introduced by the Organisation for Economic Co-Development (OECD) as part of a recommendation from 2019 meant to underpin future development of AI [7]. The United Kingdom adopted these principles for its approach to AI regulation in 2023 [8] becoming part of a government framework for future development of AI standards and guidance.

Figure 3 shows the five Trustworthy AI principles that the UK adopted as part of its framework for AI regulation. While these principles do not give a clear framework by themselves, they should nevertheless, be used to underpin any development of AI models. Recent frameworks and standards released have looked to develop the understanding of these principles into practical applications.

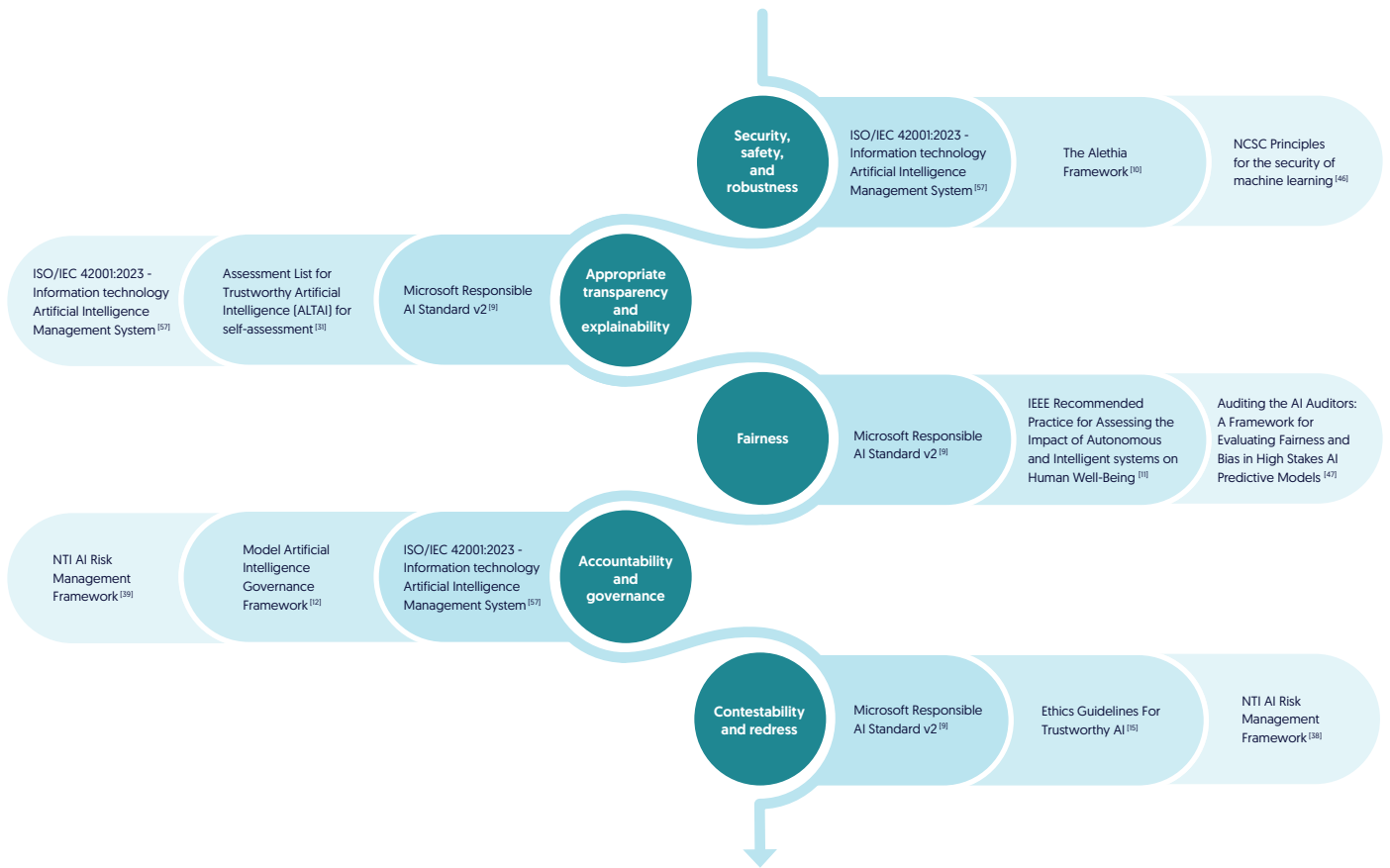



Figure 4. Diagram illustrating the pieces of reviewed literature that most comprehensively cover each trustworthy AI principle across the AI lifecycle

Existing standards and frameworks on trustworthy AI have been reviewed against the trustworthy AI principles and summarised in Figure 4. The figure highlights


the standards that cover each respective pillar most comprehensively out of 40+ reviewed standards, frameworks, policies, and white papers.

AI Assurance Techniques

Assurance techniques are tools and services that help build confidence or trust in a product, system, or even for an organisation on a wider scale, and are used regularly in industries such as finance. Applying similar assurance techniques on AI systems can help with building trust in AI and will demonstrate the trustworthiness of the AI system and consequently build and maintain trust. Similar to the trustworthy AI principles, Figure 5 highlights the standards that cover each respective assurance technique most comprehensively out of 40+ reviewed standards, frameworks, policies, and white papers. The CDEI has developed a portfolio of assurance techniques in the context of AI-enabled systems, the definitions of these different types of assurance technique can be found below:

 **Impact assessment** – Used to anticipate the effect of a system on environmental, equality, human rights, data protection, or other outcomes. Includes risk assessments and risk management strategies.

 **Compliance audit** – A review of a company's adherence to internal policies and procedures, or external regulations or legal requirements. Specialised types of compliance audit include system and process audits and regulatory inspection.

 **Certification** – A process where an independent body attests that a product, service, organisation, or individual has been tested against, and met, objective standards of quality or performance.



Bias audit – Assessing the inputs and outputs of algorithmic systems to determine if there is unfair bias in the input data, the outcome of a decision or classification made by the system.



Conformity assessment – Provides assurance that a product, service, or system being supplied meets the expectations specified or claimed, prior to it entering the market. Conformity assessment includes activities such as testing, inspection and certification.



Impact evaluation – Similar to impact assessments but are conducted after a system has been implemented in a retrospective manner.



Formal verification – Establishes whether a system satisfies some requirements using the formal methods of mathematics. These prove the correctness of an AI system's behaviour and ensure they behave as they should and do not provide incorrect outcomes.



Performance testing – Used to assess the performance of a system with respect to predetermined quantitative requirements or benchmarks to ensure it meets the performance requirements of users. Can identify bottlenecks and improvements to AI performance.



Figure 5. Diagram illustrating the pieces of reviewed literature that most comprehensively cover how to complete a particular AI assurance technique across the AI lifecycle.

Current State of Trustworthy AI

In alignment with national and international initiatives to regulate AI technology ^[8], multiple standards and frameworks have recently been developed addressing the topic of trustworthy and responsible AI. We have reviewed 40+ publications published before January 2024 including standards, frameworks, policies, and white papers, aligning them with the trustworthy AI principles and AI assurance techniques.

The statistics were collated into heatmap tables literature gaps as presented in Table 1 and Table 2, where darker cells represent intersections better covered in the literature, and lighter cells representing gaps. It is apparent that most existing standards focus the middle stages of the AI development lifecycle with the scoping & problem understanding and the retirement stages remaining poorly explored. This is a significant gap, given that many issues for AI development often stem from poor problem scoping and business problem understanding. The contestability & redress pillar is mostly relevant to the operation & monitoring stage therefore, the gap for the other stages in the AI development lifecycle is expected.

The impact and conformity assessments are well covered in the literature due to their importance. Moreover, they can be easily completed within internal development and governance teams with minimal external involvement, making them popular AI model assurance techniques. Formal certification bodies for AI models are difficult to find due to the complexity of validating standards against industry specific use cases, making it a significant gap in the AI adoption supply chain. Bias audits are important to ensure AI model fairness, a key pillar for trustworthy AI development. However, existing standards do not frequently explore the topic, highlighting the need for additional research and development in this area.

Figure 4 and Figure 5 highlight the literature that respec-

tively cover the trustworthy AI pillars and AI assurance techniques most comprehensively out of 40+ reviewed standards, frameworks, policies, and white papers. The Microsoft Responsible AI Standard, v2 ^[50], introduced in 2022, offers a comprehensive blueprint describing five crucial pillars: Accountability, Transparency, Fairness, Reliability & Safety, and Privacy & Security. Each pillar details specific goals throughout the AI development lifecycle, aiming for responsible and ethical deployment. However, while this standard provides a robust foundation, it remains very high level and requires further guidelines for its practical implementation and adaptability to diverse AI applications in manufacturing.

In tandem with Microsoft's framework, the NIST AI Risk Management Framework ^[39], also unveiled in 2022 provides a comprehensive framework that seeks to assist organisations in understanding AI system contexts and their impact on individuals and communities, urging critical assessment of potential risks and impacts. Nevertheless, challenges in implementing this framework across diverse AI contexts persist, requiring tailored approaches and seamless integration into manufacturing domains.

Moreover, the Assessment List for Trustworthy Artificial Intelligence [ALTAI] ^[5] from the European Commission, initiated in 2019, provides a self-assessment tool for AI developers grounded in fundamental ethical principles. This framework aims to foster trustworthiness by enabling self-evaluation against key ethical principles aligned with EU law. Similarly, like the AI Risk Management Framework ^[39], ALTAI accentuates the evaluation of risks and impacts of AI systems within organisations, emphasising mitigation strategies. Yet, challenges in practical implementation and adaptability across diverse AI contexts, such as manufacturing persist. This necessitates tailored strategies for different applications and wider adoption initiatives to enhance its effectiveness.

The ISO/IEC 42001:2023 - Information technology. Artificial intelligence. Management system^[57] is an international standard that outlines the requirements for developing and maintaining AI systems within an organisation. It covers most of the Trustworthy AI pillars, overlaps with the NIST Risk management Framework^[39], and discusses how AI models should integrate with other systems, which is a critical yet poorly explored topic. Although it provides guidance on how AI models should be developed to ensure safe, robust, and secure systems, there are gaps, particularly with technical and ethical themes. Additionally, as with most other documents, it lacks specifics for application to industries such as manufacturing.

Some of the existing standards in the literature effectively cover specific contexts domains such as medicine, finance, and commerce. However, the manufacturing industry requires special considerations given the variability of use cases, data streams and considerations required for the development of effective systems. Moreover, the frameworks and standards remain a paper-based exercise that explain ‘what’ should be covered; there is a limitation of practical examples on ‘how’ that can be achieved.



“Most frameworks outline ‘what’ trustworthy AI is yet lack practical explanation of ‘how’ to achieve it.”^[10]



| | Impact assessment | Compliance audit | Certification | Bias audit | Conformity assessment | Impact evaluation | Formal verification | Performance testing | Other ongoing testing |
|---------------------------------|-------------------|------------------|---------------|------------|-----------------------|-------------------|---------------------|---------------------|-----------------------|
| Scoping & problem understanding | | | | | | | | | |
| Data gathering & preparation | | | | | | | | | |
| Modelling & design | | | | | | | | | |
| Development & evaluation | | | | | | | | | |
| Deployment | | | | | | | | | |
| Live orientation & monitoring | | | | | | | | | |
| Retirement | | | | | | | | | |

Table 1. Heatmap displaying the of coverage AI assurance techniques across the AI lifecycle in the literature of AI standards and frameworks.

| | Safety, security and robustness | Appropriate transparency and explainability | Fairness | Accountability and governance | Contestability and redress |
|---------------------------------|---------------------------------|---|----------|-------------------------------|----------------------------|
| Scoping & problem understanding | | | | | |
| Data gathering & preparation | | | | | |
| Modelling & design | | | | | |
| Development & evaluation | | | | | |
| Deployment | | | | | |
| Live orientation & monitoring | | | | | |
| Retirement | | | | | |

| Legend | |
|------------------------|--|
| High coverage | |
| Medium coverage | |
| Low coverage | |
| Minimal or no coverage | |
| N/A | |

Table 2. Heatmap displaying the of coverage trustworthy AI pillars across the AI lifecycle in the literature of AI standards and frameworks.

Framework for Implementing Trustworthy AI

As can be seen from this paper, there continues to be a gap in the guidance for the practical implementation of trustworthy AI principles and assurance techniques. While the knowledge of these principles and techniques exists through a number of research papers, standards, and frameworks there are limited approaches to establishing Trustworthy AI in the manufacturing sector specifically. The approaches that are available also do not commonly inform when the different assurance techniques should be completed throughout the AI development lifecycle. In short, the frameworks outline what trustworthy AI is yet lack practical explanation of when it is necessary and how to achieve it.

Following on from the research that was gathered for this paper, the Manufacturing Technology Centre (MTC) sought to create a framework which could generate

practical applications of Trustworthy AI principles and assurance techniques for the manufacturing sector.

This framework expands on currently existing standards by providing guidance in identified gaps in current literature such as assurance assessments. It also provides users with a clear checklist at establishing trustworthy AI in their systems looking at holistically at trust in the AI development lifecycle.

The MTC's trustworthy AI framework has been developed to address the gaps in the literature. An example of the of the framework can be seen in Figure 6 and the full version is available on request. For any questions or feedback, please get in touch with the MTC's AI governance team on: aigovernance@the-mtc.org.

Conclusion

This paper demonstrates the current standards and guidance for establishing Trustworthy AI. While there is a lot of literature that is available that can be used, these are often not tailored to the manufacturing sector, nor do they provide a single clear and easy approach for AI developers to create trustworthy AI for their systems.

In this paper, the gaps to creating trustworthy AI have been identified and shows that these gaps are a restricting factor in establishing trustworthy AI and deploying trustworthy AI solutions in the future.

While this area of research continues to grow as more standards and guidance is released, there is a demand

for a unified framework to developing trust in AI systems now as the technology surrounding AI also continues to develop. The developed framework featured in this paper demonstrates one of the possible solutions to this problem in the form of a checklist framework which will enable users to understand what trustworthy AI is and when to apply principles and techniques and finally, how to demonstrate those principles and techniques for AI developers.

For the detailed Trustworthy AI framework please refer to the appendix.

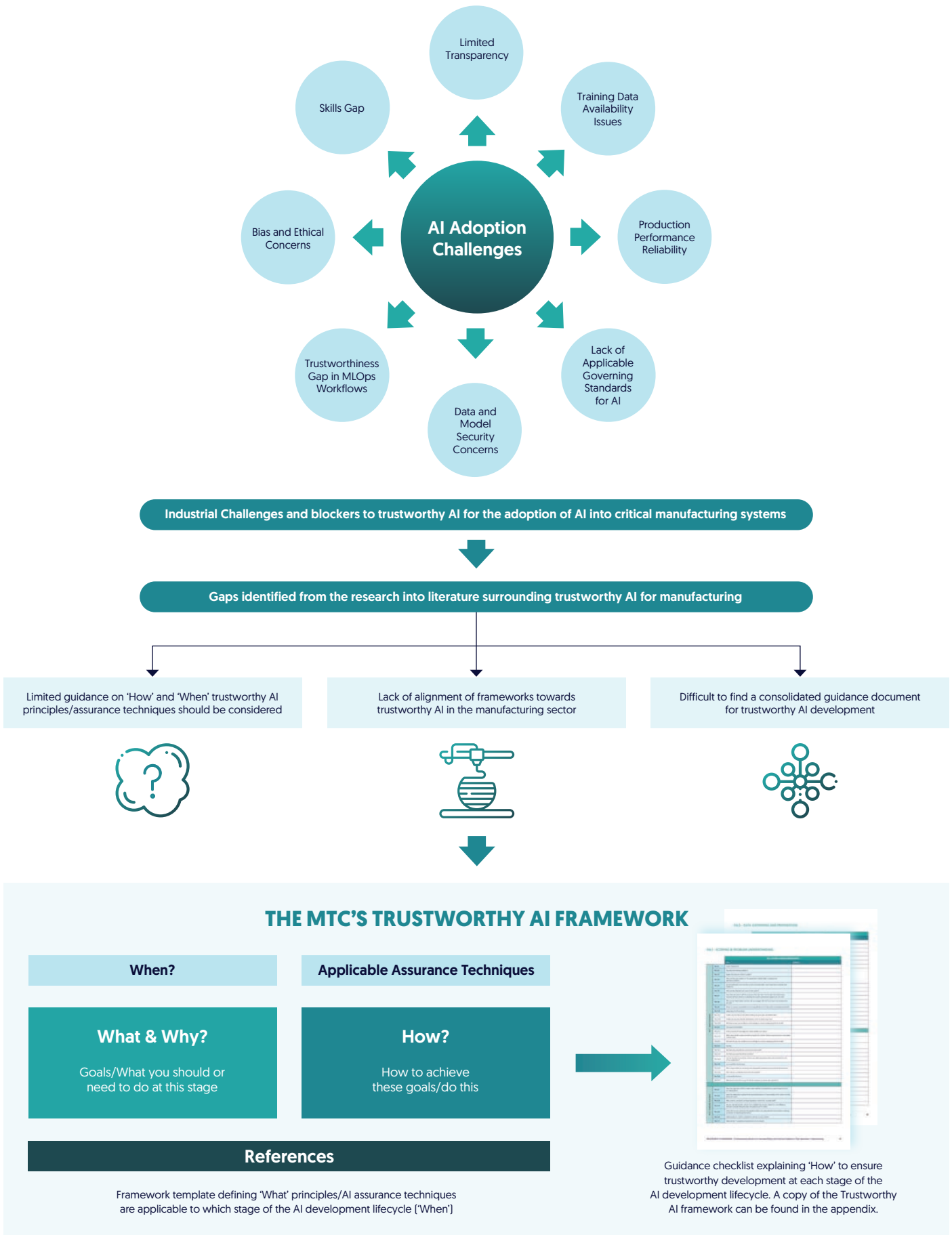


Figure 6. An outline of the gaps in the AI standard literature and example of the developed trustworthy AI framework.

References

- [1]. Evans, A. and Heimann, A. [2022] AI Activity in UK business. Available at: https://openresearch-repository.anu.edu.au/bitstream/1885/277699/1/UKAI_54.pdf [Accessed: 16 April 2024].
- [2]. Geissbauer, R., Lübben, E., Schrauf, S. and Pillsbury, S. [2018] Global Digital Operations Study 2018: Digital Champions. Available at: <https://www.pwc.com/gx/en/industries/industry-4-0.html> [Accessed: 16 April 2024].
- [3]. Bletchley Park [2023] Bletchley Park to Host AI Safety Summit. Available at: <https://bletchleypark.org.uk/bletchley-park-to-host-ai-safety-summit/> [Accessed: 16 April 2024].
- [4]. European Parliament [2023] EU AI Act: first regulation on artificial intelligence. Available at: <https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence> [Accessed: 16 April 2024].
- [5]. Hijawi, U. et al. [2023] 'Utilising Externally Hosted AI/Generative AI Services in Manufacturing: Capabilities, Limitations, Risks, Mitigations, and the role of AI Governance', Coventry: The Manufacturing Technology Centre.
- [6]. Department for Science, Innovation and Technology [2023] Portfolio of AI assurance techniques. Available at: <https://www.gov.uk/guidance/cdei-portfolio-of-ai-assurance-techniques> [Accessed: 16 April 2024].
- [7]. OECD Legal Instruments [2023] Recommendation of the Council on Artificial Intelligence. Available at: <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449> [Accessed: 16 April 2024].
- [8]. Department for Science, Innovation and Technology and Office for Artificial Intelligence [2023] AI regulation: a pro-innovation approach. Available at: <https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach> [Accessed: 16 April 2024].
- [9]. Microsoft [2022] Microsoft Responsible AI Standard, v2: General Requirements for External Release. Available at: <https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf> [Accessed: 28 March 2024].
- [10]. Rolls-Royce [2020] The Aletheia Framework. Available at: <https://www.rolls-royce.com/innovation/the-aletheia-framework.aspx> [Accessed: 28 March 2024].
- [11]. IEEE Standards Committee [2020] IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being: IEEE Standard 7010-2020. IEEE.
- [12]. Personal Data Protection Commission [PDPC] [2020] Model Artificial Intelligence Governance Framework.
- [13]. Emaminejad, N. and R. A. [2021] 'Trustworthy AI and robotics: Implications for the AEC industry', s.l.: Elsevier.
- [14]. Sujan, M., Smith-Frazer, C., Malamateniou, C., et al. [2023] 'Validation framework for the use of AI in healthcare: overview of the new British standard BS30440', *BMJ Health & Care Informatics*, 30, e100749. doi: 10.1136/bmjhci-2023-100749.
- [15]. European Union, high-level expert group on artificial intelligence [2019] Ethics guidelines for trustworthy AI. Available at: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> [Accessed: 28 March 2024].
- [16]. CIO Strategy Council [CIOOSC] [2021] Ethical design and use of automated decision systems.
- [17]. Dainow, B. and Brey, P. [2021] 'Ethics by design and ethics of use approaches for artificial intelligence', European Commission DG Research & Innovation RTD.
- [18]. Seidel, R., Schmidt, K., Thielen, N. and Franke, J. [2022]. Trustworthiness of machine learning models in manufacturing applications using the example of electronics manufacturing processes. *Procedia CIRP*, 107, pp.487-492.
- [19]. ISO/IEC [2020] ISO/IEC TR 29119-11:2020 Software and systems engineering — Software testing — Part 11: Guidelines on the testing of AI-based systems.
- [20]. Seidel, R., Schmidt, K., Thielen, N. and Franke, J. [2022] 'Trustworthiness of machine learning models in manufacturing applications using the example of electronics manufacturing processes', s.l.: Elsevier B.V.
- [21]. Frost, L., W. S. M. [2023] 'Report of TWG AI: Landscape of AI Standards', s.l.: StandICT.eu.
- [22]. OpenAI [2023] Our approach to AI safety. Available at: <https://openai.com/blog/our-approach-to-ai-safety> [Accessed: 15 May 2023].
- [23]. Personal Data Protection Commission [PDPC] [2020] Model Artificial Intelligence Governance Framework.
- [24]. IBM [n.d.] Trustworthy AI. Available at: <https://research.ibm.com/topics/trustworthy-ai> [Accessed: 28 March 2024].
- [25]. Institute of Electrical and Electronics Engineers [IEEE] Global Initiative on Ethics of Autonomous and Intelligent Systems [2019] Ethically Aligned Design [EAD]: A Vision for Prioritizing Human Wellbeing with Autonomous and Intelligent Systems.

- [26]. Institute of Electrical and Electronics Engineers (IEEE) [2022] Standard for Transparency of Autonomous Systems.
- [27]. Adedjouma, M., Adam, J.L., Akin, P., Alix, C., Baril, X., Bernard, G., Bonhomme, Y., Braunschweig, B., Cantat, L., Chale-Gongora, G. and Chihani, Z., 2022. Towards the engineering of trustworthy AI applications for critical systems-The Confiance. ai program.
- [28]. CIO Strategy Council (CIO SC) [2021] Ethical design and use of automated decision systems.
- [29]. US Association for Computing Machinery (USACM) [2017] Principles for Algorithmic Transparency and Accountability.
- [30]. Organisation for Economic Co-operation and Development (OECD) [2019] The Recommendation on Artificial Intelligence (AI).
- [31]. Ala-Pietilä, P., Bonnet, Y., Bergmann, U., Bielikova, M., Bonefeld-Dahl, C., Bauer, W., Bouarfa, L., Chatila, R., Coeckelbergh, M., Dignum, V. and Gagné, J.F. [2020] 'The assessment list for trustworthy artificial intelligence (ALTAI)', European Commission.
- [32]. The Global Partnership on Artificial Intelligence (GPAI) [2022] AI for Fair Work: AI for Fair Work Report.
- [33]. Women Leading in AI (WLiAI) [2019] White Paper: 10 Principles for Responsible AI.
- [34]. IEEE Systems, Man, and Cybernetics Society [2020] Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being.
- [35]. Bommasani, R., Klyman, K., Zhang, D., Liang, P. [2023] 'Do Foundation Model Providers Comply with the EU AI Act?', Stanford University Centre for Research on Foundation Models (CRFM).
- [36]. IEEE Robotics and Automation Society [2021] IEEE Ontological Standard for Ethically Driven Robotics and Automation Systems 7007-2021, IEEE Standards Association.
- [37]. Bird, E., F-S. N. J. R. L. E. W. a. A. W. [2020] 'The ethics of artificial intelligence: Issues and initiatives', s.l.: Panel for the Future of Science and Technology (STOA).
- [38]. Simion, M., K., C. [2023] 'Trustworthy artificial intelligence', s.l.: Asian Journal of Philosophy.
- [39]. Technology, N. I. [2023] AI Risk Management Framework, U.S Department of Commerce.
- [40]. Intelligence, I. H.-L. [2019] Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment, European Commission.
- [41]. Saif, I., B.A. [2020] 'Trustworthy AI is a Framework to Help Manage Unique Risk.', s.l.: MIT Technology Review.
- [42]. Sujan M, Smith-Frazer C, Malamateniou C, et al [2023] 'Validation framework for the use of AI in healthcare: overview of the new British standard BS30440', *BMJ Health & Care Informatics*, 30, e100749. doi: 10.1136/bmjhci-2023-100749.
- [43]. AD Hoc Committee on Artificial Intelligence (CAHAI), Policy Development Group (CAHAI-PDG) [2021] Human Rights, Democracy and Rule of Law Impact Assessment of AI systems.
- [44]. ForHumanity [2016] Independent Audit of AI Systems (IAAIS).
- [45]. Department for Science, Innovation, & Technology [2023] A pro-innovation approach to AI regulation.
- [46]. National Cyber Security Centre [2022] Principles for the security of machine learning, National Cyber Security Centre.
- [47]. Landers, R. N., S., T. [2022] 'Auditing the AI Auditors: A Framework for Evaluating Fairness and Bias in High Stakes AI Predictive Models', American Psychological Association.
- [48]. Technology, N. I. [2023] AI Risk Management Framework, U.S Department of Commerce.
- [49]. Stahl, B. C., A., J. [2023] 'A systematic review of artificial intelligence impact assessments', Springer.
- [50]. Microsoft [2022] Microsoft Responsible AI Impact Assessment Template. Available at: <https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-RAI-Impact-Assessment-Template.pdf> [Accessed: 28 March 2024].
- [51]. Raji, I. D., S.-L., A. [2020] 'Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing', Association for Computing Machinery.
- [52]. Mökander, J., F., L. [2021] 'Ethics-Based Auditing to Develop Trustworthy AI', Springer.
- [53]. Deloitte [2022] Trustworthy AI Bridging the ethics gap surrounding AI. Available at: <https://www2.deloitte.com/us/en/pages/deloitte-analytics/solutions/ethics-of-ai-framework.html> [Accessed: 28 March 2024].
- [54]. FIMA, D. H. [2023] Artificial Intelligence/Machine Learning Policy for SMEs, Institute of Mathematics.
- [55]. Floridi, L., H., M. [2022] 'A procedure for conducting conformity assessment of AI systems in line with the EU Artificial Intelligence Act', University of Oxford.
- [56]. Urban, C., M., A. [2021] 'A Review of Formal Methods', Cornell University.
- [57]. ISO/IEC [2023] ISO/IEC 42001:2023 - Information technology Artificial intelligence Management system.

Appendix

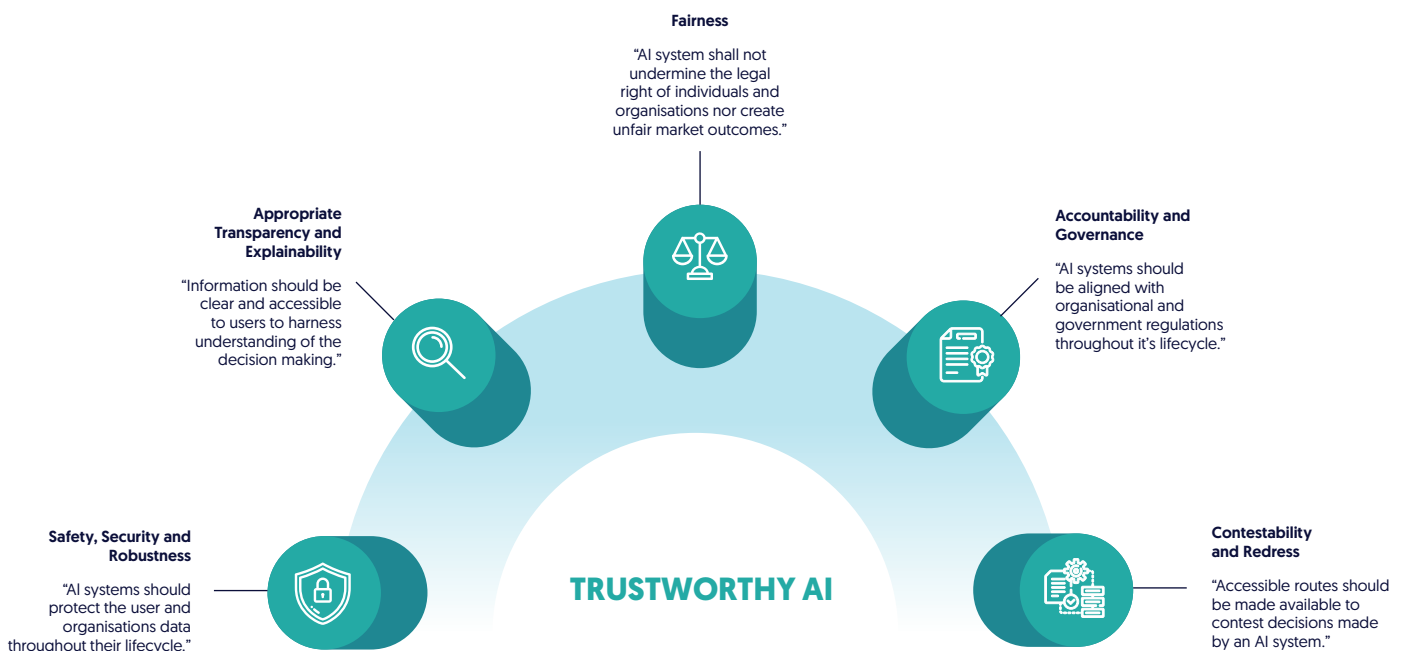
Trustworthy AI Framework for Industry Applications

WHAT IS TRUSTWORTHY AI?

Trustworthy AI refers to the development and deployment of AI systems that are secure, robust, transparent, fair, and aligned with governance values. With the growing implementation of AI by companies in many industries looking to take advantage of the potential benefits of AI, it is vital that there are strong ethical considerations made throughout the AI lifecycle to ensure that these systems operate fairly and with minimal negative impact. This framework acts as practical guidance across the AI lifecycle that will demonstrate how to achieve trustworthy AI in an industrial setting/application.

TRUSTWORTHY AI PRINCIPLES

Trustworthy AI often encompasses a set of principles and techniques aimed at ensuring that AI technology is developed and operated in a manner that inspires confidence by the user, developer, and all associated stakeholders. Below are the set of principles used to guide this framework that align with those set out by the UK Department for Science, Innovation and Technology (DSTI) and Centre for Data Ethics and Innovation (CDEI) ^{[1], [2]}:



1. A pro-innovation approach to AI regulation - GOV.UK (www.gov.uk)
2. Portfolio of AI assurance techniques - GOV.UK (www.gov.uk)

AI ASSURANCE TECHNIQUES

Assurance techniques are tools and services that help build confidence or trust in a product, system, or even for an organisation on a wider scale, and are used regularly in industries such as finance. Applying similar assurance techniques on AI systems can help with building trust in AI and will demonstrate the trustworthiness of the AI system and consequently build and maintain trust. Similar to the trustworthy AI principles, Figure 5 highlights the standards that cover each respective assurance technique most comprehensively out of 40+ reviewed standards, frameworks, policies, and white papers. The CDEI has developed a portfolio of assurance techniques in the context of AI-enabled systems, the definitions of these different types of assurance technique can be found below:



Impact assessment – Used to anticipate the effect of a system on environmental, equality, human rights, data protection, or other outcomes. Includes risk assessments and risk management strategies.



Compliance audit – A review of a company's adherence to internal policies and procedures, or external regulations or legal requirements. Specialised types of compliance audit include system and process audits and regulatory inspection.



Certification – A process where an independent body attests that a product, service, organisation, or individual has been tested against, and met, objective standards of quality or performance.



Bias audit – Assessing the inputs and outputs of algorithmic systems to determine if there is unfair bias in the input data, the outcome of a decision or classification made by the system.



Conformity assessment – Provides assurance that a product, service, or system being supplied meets the expectations specified or claimed, prior to it entering the market. Conformity assessment includes activities such as testing, inspection and certification.



Impact evaluation – Similar to impact assessments but are conducted after a system has been implemented in a retrospective manner.

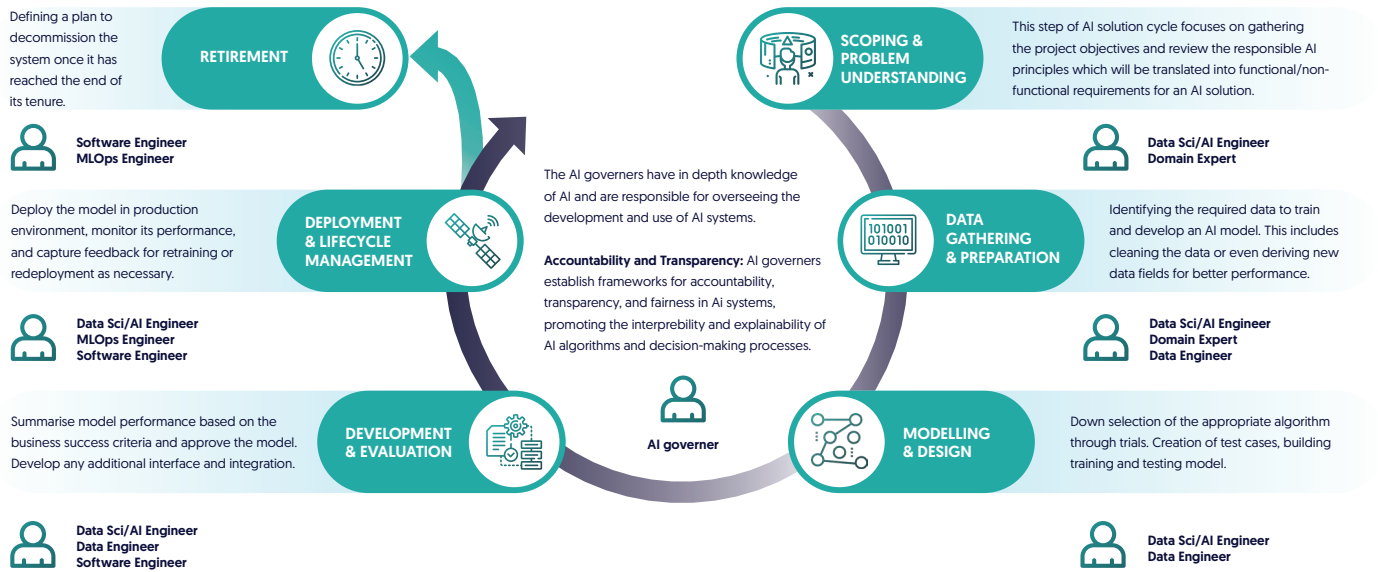


Formal verification – Establishes whether a system satisfies some requirements using the formal methods of mathematics. These prove the correctness of an AI system's behaviour and ensure they behave as they should and do not provide incorrect outcomes.



Performance testing – Used to assess the performance of a system with respect to predetermined quantitative requirements or benchmarks to ensure it meets the performance requirements of users. Can identify bottlenecks and improvements to AI performance.

AI DEVELOPMENT LIFECYCLE



| TAI.1 - SCOPING & PROBLEM UNDERSTANDING | | | |
|---|---|---|--|
| | Item | Evidence | |
| TAI.1.1 - Impact Assessment | TAI.1.1.1 | Impact Assessment | |
| | TAI.1.1.2 | Describe the business problem? | |
| | TAI.1.1.3 | Explain the purpose of the AI system? | |
| | TAI.1.1.4 | Why is AI the best solution for this application? Explain while comparing it to alternative solutions. | |
| | TAI.1.1.5 | List all stakeholders and how the model could potentially impact them both positively and negatively. | |
| | TAI.1.1.6 | What are the intended use-cases for this system? | |
| | TAI.1.1.7 | How does the system fulfil the purpose of the use-case? List the main Key Performance Indicators (KPIs) to allow for evaluating the model's performance against the use case? | |
| | TAI.1.1.8 | Who are the stakeholders and how will you engage with them to ensure their requirements are met? | |
| | TAI.1.1.9 | Person or persons responsible for ensuring adherence to Trustworthy AI principles/standards? | |
| | TAI.1.1.10 | Safety/Security/Robustness: | |
| | TAI.1.1.10.1 | Outline security risks for the data including any personally identifiable data? | |
| | TAI.1.1.10.2 | Outline any security risks the introduction of the AI model may incur? | |
| | TAI.1.1.10.3 | Will there be any loss of skillsets or knowledge as a result of deploying this AI model? | |
| | TAI.1.1.11 | Transparency/Reliability: | |
| | TAI.1.1.11.1 | Is the proposed AI type/algorithm been detailed out clearly? | |
| | TAI.1.1.11.2 | What value will this system provide compared to a human? (if the proposed system automated a human task.) | |
| | TAI.1.1.11.3 | Will there be any loss of skillsets or knowledge as a result of deploying this AI model? | |
| | TAI.1.1.12 | Fairness: | |
| | TAI.1.1.12.1 | Are there any potential bias concerns for the model? | |
| | TAI.1.1.12.2 | Are there any potential ethical concerns? | |
| | TAI.1.1.12.3 | Can the developed model be used for any other application which may be harmful to any of the stakeholders? | |
| | TAI.1.1.13 | Accountability/Governance: | |
| | TAI.1.1.13.1 | Who is responsible for overseeing and assuring the development process (AI Governance)? | |
| | TAI.1.1.13.2 | How will any confidential information be handled? | |
| TAI.1.1.14 | Contestability/Redress: | | |
| TAI.1.1.14.1 | What level of Human-in-Loop HIL will be necessary to ensure safe operation? | | |
| TAI.1.2 - Conformity Assessment | TAI.1.2.1 | Have the objectives of the AI system been defined, documented and gained approval from the stakeholders? | |
| | TAI.1.2.2 | Have the stakeholders agreed to the appointed persons of responsibility of the system and the listed use-cases? | |
| | TAI.1.2.3 | What policies, standards and legal regulations need to be complied with? | |
| | TAI.1.2.4 | Has any unit testing been carried out to establish the success criteria? (i.e., the difference between a human doing this task compared to an AI model.) | |
| | TAI.1.2.5 | Define the success criteria for the planned solution including desired improvement to existing processes and desired performance? | |
| | TAI.1.2.6 | Define testing to conform compliance with the success criteria? | |
| | TAI.1.2.7 | What are the IT compliance requirements for the model? | |

TAI.2 - DATA GATHERING AND PREPARATION

| TAI.2 - DATA GATHERING & PREPERATION | | | |
|--------------------------------------|---|---|--|
| | Item | Evidence | |
| TAI.2.1 - Impact Assessment | TAI.2.1.1 | List all potential data sources accounting for: | |
| | TAI.2.1.1.1 | Data sensitivity and export sensitivity classification labels? | |
| | TAI.2.1.1.2 | Any potential risks for dual use? | |
| | TAI.2.1.1.3 | Is the data personally identifiable? | |
| | TAI.2.1.1.4 | If so, is the data owner aware of the data usage in compliance with data protection acts? | |
| | TAI.2.1.1.5 | Methodology and architecture for data storage and processing with adequate security? | |
| | TAI.2.1.2 | Complete exploratory data analysis (EDA) highlighting: | |
| | TAI.2.1.2.1 | The data type: <ul style="list-style-type: none"> ▶ Images ▶ Text ▶ Numbers ▶ Time Series ▶ Other | |
| | TAI.2.1.2.2 | The balance of the data targets? (if applicable) | |
| | TAI.2.1.2.3 | All potential bias concerns with the collected data? | |
| | TAI.2.1.2.4 | All potential breach of policies or regulations (e.g. GDPR) with collected data? | |
| | TAI.2.1.2.5 | What is the size of the dataset? | |
| | TAI.2.1.2.6 | Data gaps detailing the process of identifying them? | |
| | TAI.2.1.2.7 | Is the collected data representative of the real-life data distribution? | |
| | TAI.2.1.2.8 | Summary data statistics? | |
| TAI.2.1.2.9 | Any other notable discoveries? | | |
| TAI.2.1.3 | Outline the data processing steps required to prepare the data AI model development. | | |
| TAI.2.1.4 | Will any data augmentation be required? If so, please outline the methodology. | | |
| TAI.2.1.5 | What legal compliance is required to collect this data and how will you ensure compliance? | | |
| TAI.2.2 - Conformity Assessment | TAI.2.2.1 | What legal compliance is required to collect this data and how will you ensure compliance? | |
| | TAI.2.2.2 | Has a data summary datasheet been provided? (This includes details on data provenance, data collection methods, and data preprocessing steps undertaken.) | |
| | TAI.2.2.3 | Are the data processing pipelines in compliance with the relevant policies, standards and regulations highlighted in the 'Scoping' stage? | |
| | TAI.2.2.4 | Is a data management and security process in place? | |
| | TAI.2.2.5 | Is the data processing environment secure? | |
| | TAI.2.2.6 | Is the environment on-premises or cloud? | |
| | TAI.2.2.7 | If cloud: | |
| | TAI 2.2.7.1 | Who are the cloud providers? | |
| | TAI 2.2.7.2 | Where are the geographical locations of the servers? | |
| | TAI 2.2.7.3 | Has a review of the cloud provider's security certification been reviewed? | |
| | TAI 2.2.7.4 | Has a review of the cloud provider's compute and data residency compliance been reviewed? | |
| | TAI 2.2.7.5 | What communication protocols and safety measures and in place to ensure secure access and processing? | |
| | TAI 2.2.8 | If on-premise: | |
| | TAI 2.2.8.1 | What is the full system architecture? | |
| TAI.2.2.8.2 | What communication protocols and safety measures and in place to ensure secure access and processing? | | |
| TAI 2.2.9 | Has a risk register of the use of the collected data been provided? | | |

TAI.3 - MODELLING AND DESIGN

| TAI.3 - MODELLING & DESIGN | | | |
|---------------------------------|-------------|---|--|
| | Item | Evidence | |
| TAI.3.1 - Impact Assessment | TAI.3.1.1 | Detail the down-selection process to identify the suitable model/model-architecture? | |
| | TAI.3.1.2 | Provide a detailed description of the model algorithm with supporting architecture diagrams. | |
| | TAI.3.1.3 | Is this a pre-trained foundational model? | |
| | TAI.3.1.4 | If so, please detail any potentially pre-existing bias in the model and how you plan to mitigate for it. | |
| | TAI.3.1.5 | The chosen model's Technology Readiness Level (TRL)? | |
| | TAI.3.1.6 | What feature scaling techniques have been used and why? | |
| | TAI.3.1.7 | Is the chosen model expected to manage tasks of high complexity? | |
| | TAI.3.1.8 | Provide the suggested architecture for the model development, testing and production environments. | |
| | TAI.3.1.9 | What is the expected carbon impact of training and running this model? | |
| | TAI.3.1.10 | How will the model interface/integrate with the relevant stakeholders? | |
| | TAI.3.1.11 | Can the model be used for malicious applications? | |
| | TAI.3.1.12 | Detail what model security considerations will be implemented. | |
| TAI.3.2 - Conformity Assessment | TAI.3.1.1 | Has the algorithm been reviewed and approved for use, this includes checking licenses and permissions for commercial use if applicable. | |
| | TAI.3.1.2 | Justification over whether to use locally developed model or an off-the shelf model. | |
| | TAI.3.1.3 | Justification if the model is deployed on-prem or on an external server and interacted with via API. | |
| | TAI.3.1.3.1 | Hosting on an external server reduces the effort to manage hardware but raises concerns over safety/security. | |
| | TAI.3.1.3.2 | Local hosting has computationally expensive overheads and management. | |
| | TAI.3.1.4 | Has a strategy been defined to validate this model? | |
| | TAI.3.1.5 | Are the stakeholders satisfied with the integration of the model with the system application? | |
| | TAI.3.1.6 | How will the model security be assured and how will you protect against any information breach through the model? | |

[Continued on next page]

| | | | |
|-------------------------------|---|---|--|
| TAI.3.3 - Performance Testing | TAI.3.3.1 | Detailed plan on how the model will be tested during the design and development stages as well as the success criteria for the model. | |
| | TAI.3.3.2 | Define unit tests, outlining the success and failure criteria for the model: | |
| | TAI.3.3.2.1 | Accuracy and robustness of the model. | |
| | TAI.3.3.2.2 | Model security. | |
| | TAI.3.3.2.3 | Bias and fairness testing. | |
| | TAI.3.3.2.4 | Testing the model boundaries and identifying limitations. | |
| | TAI.3.3.2.5 | Testing model transparency and explainability. | |
| | TAI.3.3.2.6 | What is the expected carbon impact of the model? | |
| | TAI.3.3.3 | Specify the performance tests and expected success results: | |
| | TAI.3.3.3.1 | Supervised: <ul style="list-style-type: none"> ▶ Accuracy ▶ Precision ▶ Recall ▶ F1 score ▶ Area under the curve [AUC] ▶ Loss (please specify) ▶ Other | |
| | TAI.3.3.3.2 | Unsupervised: <ul style="list-style-type: none"> ▶ Silhouette coefficient ▶ Calinski-Harabasz index ▶ Davies-Buldin index ▶ Other | |
| | TAI.3.3.3.3 | Natural Language Processing (including LLMs): <ul style="list-style-type: none"> ▶ Bilingual Evaluation Understudy (BLUE) ▶ Recall-Oriented Understudy for Gisting Evaluation (ROGUE) ▶ Other | |
| TAI.3.3.3.4 | Other: <ul style="list-style-type: none"> ▶ Please specify | | |
| TAI.3.3.4 | Establish lines of responsibility for who will be testing the model and what tests they will be conducting. | | |

TAI.4 - DEVELOPMENT

| TAI.4 - DEVELOPMENT | | | |
|---------------------------------|--|---|--|
| | Item | Evidence | |
| TAI.4.1 - Impact Assessment | TAI.4.1.1 | Have any changes been made to the planned deployment infrastructure? Explain with justification. | |
| | TAI.4.1.2 | Highlight and explain any changes to the planned model and development/ | |
| | TAI.4.1.3 | If the model is intended for internal use only, how easily available will this system be to the business? | |
| | TAI.4.1.3.1 | Will it only be available to authorised departments? | |
| | TAI.4.1.3.2 | Or will it be made available to everyone in the business? | |
| | TAI.4.1.4 | If the model is intended for commercial use, how will the user access and interact with the system? | |
| | TAI.4.1.4.1 | Will the system have tiered access [i.e. free/premium tier etc.]? Will the systems full functionality be available across these tiers? | |
| | TAI.4.1.5 | Outline any planned training and upskilling required for stakeholders/users prior to model deployment. | |
| | TAI.4.1.6 | Detail the planned continuous monitoring process post-deployment. | |
| | TAI.4.1.7 | Detail and document all development processes and findings. | |
| TAI.4.1.8 | Can the overall system be manipulated or altered for malicious applications? | | |
| TAI.4.1.9 | Detail what model security considerations will be implemented. | | |
| TAI.4.2 - Conformity Assessment | TAI.4.2.1 | Are the defined unit tests successful? | |
| | TAI.4.2.2 | Are the required performance expectations met? | |
| | TAI.4.2.3 | Have disclaimers been added to alert users that an AI model is in use with this system prior to deployment? | |
| | TAI.4.2.4 | If using personal data, are disclaimers provided before a user submits their personal data? | |
| | TAI.4.2.5 | Have the appropriate persons been identified to oversee the accountability and governance of the system? | |
| | TAI.4.2.6 | Is there an agreed upon productionisation plan with the relevant stakeholders? | |
| | TAI.4.2.7 | Is the production-ready version any different to how the system functions in contrast to what was initially agreed upon in the scoping stage? | |
| TAI.4.3 - Performance Testing | TAI.4.3.1 | Identify or highlight technical limitations of the system. | |
| | TAI.4.3.2 | Detail the results of the performance checks. | |
| | TAI.4.3.3 | Provide a comprehensive overview of the unit tests, their outcomes, and any deviations from expected results. Additionally, describe how any failed tests will be addressed and resolved. | |
| | TAI.4.3.4 | Are the tests providing relevant and useful information about the development of the model? | |
| | TAI.4.3.5 | Is the testing timeframe still consistent with the project plan? | |
| | TAI.4.3.6 | Are any additional required tests being identified? | |
| | TAI.4.3.7 | Address any performance discrepancies between the production-ready version and the initially agreed-upon specifications from the modelling & design stage. | |

TAI.5 - DEPLOYMENT & LIFECYCLE MANAGEMENT

| TAI.5 - DEPLOYMENT & LCM | | | |
|---------------------------------|-------------|---|--|
| | Item | Evidence | |
| TAI.5.1 - Impact Evaluation | TAI.5.1.1 | Are there any instances of anomalous behaviour by the system that wasn't predicted before deployment? If so, include strategies for anomaly detection and how these instances were addressed or mitigated. | |
| | TAI.5.1.2 | Has the system been recalled since deployment due to behaviour drift or performance falling below the threshold? | |
| | TAI.5.1.3 | What are system's impacts on the process in which it is integrated with? Please state all positive and negative influences. | |
| | TAI.5.1.4 | What have been the ethical and societal impacts of deploying this model? | |
| | TAI.5.1.5 | How has the system impacted its stakeholder? Please state all positive and negative influences. | |
| | TAI.5.1.6 | Evaluate the realised value of the system after deployment by comparing it to the KPIs and requirements specified in the scoping section. | |
| TAI.5.2 - Conformity Assessment | TAI.5.2.1 | Detail the redress plan for an outdated training set, outlining steps to update or retrain the model and its potential impact on system performance. Has the system been effectively handed over to the MLOps team? | |
| | TAI.5.2.2 | Specify the handover process to the MLOps team, ensuring clarity on responsibilities, ongoing monitoring & maintenance tasks. | |
| | TAI.5.2.3 | Are the stakeholders satisfied with the overall delivery of the system? | |
| | TAI.5.2.4 | Is the deployment compliant with cyber-security protocols? | |
| | TAI.5.2.5 | Is the deployment compliant with data protection protocols? | |
| | TAI.5.2.6 | Are satisfactory AI security measures being taken to protect the model and its accompanying data? | |
| | TAI.5.2.7 | Outline your model versioning and rollback policies. | |
| TAI.5.3 - Performance Testing | TAI.5.3.1 | Identify or highlight technical limitations of the system. | |
| | TAI.5.3.2 | Detail the results of the performance checks. | |
| | TAI.5.3.2.1 | Data drift. | |
| | TAI.5.3.2.2 | Model security. | |
| | TAI.5.3.2.3 | Bias monitoring. | |
| | TAI.5.3.2.4 | Health and operation monitoring. | |
| | TAI.5.3.2.5 | Exploitability effectiveness. | |
| | TAI.5.3.2.6 | Other, please specify. | |
| | TAI.5.3.3 | Is there going to be any automatic validation metrics built into the model to show current effectiveness during lifecycle? | |
| | TAI.5.3.4 | Will there be an active user feedback system in place. How will the feedback be used in evaluating the performance of the model? | |

TAI.6 - RETIREMENT

| TAI.6 - RETIREMENT | | | |
|-----------------------------|-------------|--|--|
| | Item | Evidence | |
| TAI.6.1 - Impact Evaluation | TAI.6.1.1 | Did the system meet its functional obligations during its operational lifetime? | |
| | TAI.6.1.2 | Will this system be replaced by a newer model or is an alternative technology being sought? | |
| | TAI.6.1.3 | Was the forecasted value added from this system realised? | |
| | TAI.6.1.4 | Please detail the reasons for retiring the system. | |
| | TAI.6.1.5 | Evaluate the realised value from the system by comparing the forecasted benefits or objectives with the actual outcomes achieved. | |
| | TAI.6.1.6 | Evaluate the impact the system had on the following: | |
| | TAI.6.1.6.1 | Stakeholders including user feedback. | |
| | TAI.6.1.6.2 | Integrated process and overall effect on operation. | |
| | TAI.6.1.6.3 | Ethical and societal impacts. | |
| | TAI.6.1.6.4 | Environmental and resource impact. | |
| | TAI.6.1.7 | Address any potential risks or dependencies associated with retiring the system, such as ensuring a smooth transition without disrupting ongoing operations or services that rely on the retired system. | |

mtc
Manufacturing
Technology Centre

CATAPULT
High Value Manufacturing

the-mtc.org