Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference
Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage, Marja Ylönen
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P7834-cd

Towards AI trustworthiness assessment framework for railway applications

Asma Ladj

Technological Research Institute Railenium, 180 rue Joseph-Louis Lagrange, Valenciennes, F-59300, France
E-mail: asma.ladj@railenium.eu

Abderraouf Boussif,

Univ. Gustave Eiffel, COSYS-ESTAS, 20 rue Élisée Reclus, Villeneuve d'Ascq, F-59650, France
E-mail: abderraouf.boussif@univ-eiffel.fr

While artificial intelligence (AI) has the potential to significantly enhance performance of railway transportation and mobility, ensuring its trustworthiness and safety remains a serious challenge. Indeed, the deployment of AI systems in general, and particularly in railway applications, gives rise to multidisciplinary concerns spanning ethical, social, economic, and technical dimensions. This paper first presents an overview of the key concepts related to AI trustworthiness assessment and then outlines the foundational steps for establishing a framework to assess AI trustworthiness in the railway sector, in light of the EU AI Act. Specifically, it explores the parallels between railway risk assessment and AI trustworthiness assessment, while adapting the definition of risk to encompass the AI-related risks, and extending the analysis activities to consider additional trustworthiness attributes.

Keywords: Artificial Intelligence (AI), AI Systems, Railway applications, AI trustworthiness assessment, AI Act

1. Introduction

The railway sector has witnessed a growing interest in integrating artificial intelligence (AI) technologies across various domain applications, including predictive maintenance and inspection, automated driving, traffic management, and planning (Tang et al., 2022). The adoption of AI-driven solutions aims to significantly improve operational efficiency, safety, and reliability, while simultaneously reducing downtime and operational costs.

The railway industry is conventionally focusing on developing and integrating dependable software and systems with predictable, tractable and verifiable behavior. Indeed, the railway sector is recognized as safe and a highly regulated domain. To maintain this high level of safety, a safety management system, along with established standards and guidelines for specification and demonstration of safety and dependability, is in place to evaluate the impact of any new application or technology introduced into the system. Consequently, AI-driven technologies and applications are no exception to the rule (Donato et al., 2022).

While in traditional (non-AI) systems, the primary focus is on safety, dependability, and security, AI systems introduce additional technical, societal, and ethical dimensions that must be rigorously evaluated (Wing, 2021). Indeed, due to their black-box nature, complexity and autonomy, AI systems often exhibit unpredictable behavior, leading to critical safety concerns related to uncontrollability, ethical conflicts, and long-term socioeconomic impacts. These challenges underscore the need to consider additional system properties and factors such as transparency, explainability, accountability, and the ethical implications of decision-making, particularly when deployed in safety-critical applications (Macrae, 2024). All these factors and attributes form the foundation of what is known as trustworthiness of AI (Awadid et al., 2024; Mattioli et al., 2024).

In this context, global and European efforts are playing a key role in providing technical standards and regulatory frameworks for the development and use of advanced AI systems (Jeon, 2024). The overall objective is to mitigate both material (e.g., safety and health of individuals, damage to property) and immaterial (e.g., loss of privacy, limitations to human rights, human dignity, discrimination) harms associated with AI systems. The European Commission aims to a harmonized regulatory framework for AI development and deployment, through the so-called European Union Artificial Intelligence Act (EU AI Act) (European Commission, 2024).

Aiming to establish a framework for assessing AI trustworthiness in the railway sector, this paper presents an overview of the key concepts related to AI trustworthiness assessment, as well as the foundational steps, structured in terms of hierarchical processes and activities, toward its establishment. Specifically, we investigate the parallels between railway risk assessment and AI trustworthiness assessment, while adapting the definition of risk to encompass the AI-related risks, and extending the analysis activities to consider additional trustworthiness attributes.

The remainder of the paper is organized as follows. Section 2 discusses definitions of AI, AI systems and enumerates AI applications in the railway sector. Section 3 focuses on AI trustworthiness concept and reports relevant EU guidelines, regulations, and standards for AI. In section 4, the safety assessment process is recalled and then, its extension to AI trustworthiness is discussed. Finally, Section 5 brings some concluding remarks.

2. AI systems and railway applications

We firstly present the concept of AI systems and then discuss how AI technologies are applied and deployed within the railway industry.

2.1. *AI systems*

There is no universally accepted definition of artificial intelligence (AI). As a scientific discipline, the ISO/IEC 2382:2015 standard[a] defines AI as "*a branch of computer science devoted to developing data processing systems that perform functions normally associated with human intelligence, such as reasoning, learning, and self-improvement*". ISO/IEC TR 24028:2020[b] provides a more technical definition, as the "*capability of an engineered system to acquire, process, and apply knowledge and skills*". Regarding AI systems, the High-Level Expert Group On Artificial Intelligence (AI HLEG)[c] defines them as "*software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal*".

For the purpose of AI Act Regulation, a legal definition of AI system is " *machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments*" (European Commission, 2024).

All definitions technically agree that AI systems can be integrated either entirely in software (e.g., image analysis tools), or embedded in hardware devices (e.g., self-driving cars). They also converge on the main capabilities of AI systems: environment perception, information processing, decision-making, and actuation (Stettinger et al., 2024).

Generally, three classes of AI approaches can be distinguished, (1) *machine learning* approaches (also known as data-driven AI, including supervised, unsupervised and reinforcement learning), (2) *logic- and knowledge-based* approaches (also known as symbolic AI, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems), and (3) *statistical* approaches (including Bayesian estimation, search and optimization methods).

[a] ISO/IEC 2382:2015 Information technology - Vocabulary.
[b] ISO/IEC TR 24028:2020 Information technology - Artificial intelligence - Overview of trustworthiness in artificial intelligence.

[c] https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56341

2.2. *AI Applications in Railway*

While AI integration is still emerging within the railway sector, it demonstrates significant potential to overcome key challenges. Structural railway subsystems, including rolling stock, infrastructure, energy, and trackside/on-board control-command and signaling, can benefit from AI-driven solutions to enhance their efficiency, reliability, and safety. Similarly, functional subsystems, such as operation and traffic management, maintenance, and telematics services applications, present opportunities for AI to reduce costs and improve service quality.

A wide spectrum of railway applications is being covered by AI, including Maintenance and Inspection, Traffic Planning and Management, Safety and Security, Autonomous Driving and Control, Transport Policy, Passenger Mobility, and Business & Finance. The great majority of research has focused on Maintenance and Inspection tasks, followed by Traffic Planning and Management, while sub-fields of Safety and Security, Autonomous Driving and Control, and Passenger Mobility received only limited attention (Tang et al., 2022).

This imbalanced focus can be attributed to the current shift toward predictive maintenance, characterized by a demonstrable reduction in operational costs, fostering increased industry adoption and stimulating further research endeavors. Moreover, the widespread use of sensors has led to an abundance of data, enabling the development of advanced faults/defects diagnosis and prediction methods. On the other hand, the other subdomains face greater technical, regulatory, and societal challenges. Safety and Security related applications require rigorous testing, validation, and regulatory approval before deploying AI solutions. Autonomous driving and control require complex decision-making algorithms, efficient integration with existing systems, and compliance with strict safety standards. Passenger mobility applications are highly dependent on human behavior, preferences, and dynamic urban contexts, which are more difficult to model and optimize with AI.

3. AI trustworthiness

This section focuses on AI trustworthiness. First, we discuss the definitions and key characteristics of trustworthiness. Next, we succinctly present the regulatory and normative frameworks governing AI trustworthiness. Finally, we examine the main efforts and directions for managing and assessing AI systems' trustworthiness.

3.1. *Trustworthiness concept*

Driven by the impressive performance of AI systems, AI technologies continue to transform various domains. However, as AI advances, stakeholders have increasingly recognized the need to address its inherent risks. For instance, AI-based image analysis tools (for medical diagnostics, self-driving car perception, etc.) may misclassify images when noise is added. Bias in AI-driven decision-making has also been extensively observed, such as with corporate recruiting algorithms exhibiting bias against women. These challenges highlight that AI-related risks are not merely technical concerns but fundamental societal issues.

According to ISO/IEC TR 24028:2020, trustworthiness is "*ability to meet stakeholders' expectations in a verifiable way*". It was noted that trustworthiness depends on the context, sector, product/service, data, and technology. Similarly to dependability, trustworthiness is a meta-term which combines several aspects in a quite generic way. The set of trustworthiness properties for AI systems, in contrast to traditional computing systems, needs to be extended beyond reliability, security, privacy, and usability to include properties such as fairness, robustness, accountability, and explainability (Wing, 2021).

The Ethics Guidelines for Trustworthy Artificial Intelligence published by the AI HLEG[d] , states that trustworthy AI should be (1) *lawful* (complying with all applicable laws and regulations), (2) *ethical* (adherence with ethical principles and values), and (3) *robust* (both from a technical perspective while taking into account its so-

[d] https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

cial environment). The AI HLEG also defined four ethical principles: respect for human autonomy, prevention of harm, fairness, and explainability. Based on that, seven key requirements, applying to all stakeholders (developers, deployers, end-users, and society at large), should be satisfied throughout the entire life cycle of AI systems: Human Agency and Oversight, Technical Robustness and Safety, Privacy and Data Governance, Transparency, Diversity, Non-discrimination and Fairness, Societal and Environmental Well-being, Accountability. Each requirement is crucial, yet none alone is sufficient to achieve trustworthy AI (Stettinger et al., 2024). An Assessment List for Trustworthy Artificial Intelligence (ALTAI) is defined by the AI HLEG[e], to provide a structured framework for organizations to self-assess their AI systems during development, deployment, procurement, or use. Its adaptable design allows organizations to tailor ALTAI to their specific sector and needs.

3.2. *AI regulation framework*

As the existing legislation proves insufficient to address the risks posed by AI technologies, AI regulation has emerged as a crucial policy issue worldwide. The European Union (EU) plays an important role in shaping the future of AI regulation. Indeed, the European Commission's approach to AI regulation has undergone substantial development, from the foundational blueprint set out in the White Paper on Artificial Intelligence (European Commission, 2020) to the more refined and structured regulatory framework, the so-called Artificial Intelligence Act (AI Act) (European Commission, 2024). The proposed EU AI Act aims to create a unified internal market for trustworthy AI by establishing a harmonized legal framework. The key objectives are ensuring AI safety and EU law compliance, fostering investment and innovation through legal certainty, and strengthening governance and enforcement of fundamental rights and safety.

[e]The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self Assessment, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342

The main characteristic of the AI Act is its risk-based approach, focusing on the potential harm to health, safety, and fundamental human rights. In the AI Act regulatory framework, requirements and obligations for the development, market placement, and use of AI systems, are tailored to the level and scope of potential risks. Four levels are identified: (*i*) low or minimal risk (no regulatory restrictions), (*ii*) limited risk (transparency obligations), (*iii*) high risk (stringent regulatory requirements), and (*iv*) unacceptable risk (prohibited AI practices).

3.3. *AI trustworthiness standardization*

In recent years, several expert working groups and standardization committees have been actively developing concepts, guidelines, best practices, methods, and tools for managing and assessing the trustworthiness of AI systems, particularly in safety-critical domains.

In addition to terminology and concepts introduced in (*ISO/IEC 22989:2022*) and (*ISO/IEC 23053:2022*), an overview of trustworthiness in AI is proposed in (*ISO/IEC TR 24028:2020*), and a framework for the development and trustworthiness of autonomous/cognitive systems is established in (*VDE-AR-E 2842-61: 2021*) Additionally, standards such as (*ISO/IEC 42001:2024*) for AI management systems, (*ISO/IEC 23894:2023*) for risk management, and (*ISO/IEC 25059:2023*) for AI system quality models offer frameworks for organizations to ensure reliable and responsible AI deployment.

To efficiently develop AI systems and data models, several standards have been established, including (*DIN SPEC 92001:2019*), (*ISO/IEC 5338:2023*), (*ISO/IEC 5339:2024*), and (*ISO/IEC 8183:2023*). These standards provide comprehensive guidelines for AI development, deployment, and data governance. Notice that most AI-related standards have focused on key attributes and quality characteristics of trustworthiness, including safety (*ISO/IEC TR 5469:2024*), bias (*ISO/IEC TR 24027:2021*), robustness (*ISO/IEC 24029 series*), transparency (*ISO/IEC DIS 12792*), and explainability (*ISO/IEC CD TS 6254*).

For a comprehensive review of standardization efforts related to AI trustworthiness, readers can refer to (Jeon, 2024).

3.4. *AI trustworthiness management*

Aware of the sociotechnical hazards and risks arising from the deployment of AI and learning-based systems, considerable effort has been devoted to ensuring the safety and trustworthiness of AI systems (NIST, 2023). Globally, three primary questions are considered: (*i*) the development of organizational-level trustworthy management frameworks, (*ii*) the establishment of trustworthy assessment processes, and (*iii*) the identification and categorization of AI-related hazards and risks.

The standard (ISO/IEC 42001:2024) is considered as a main pillar for AI management in organizations. It specifies the requirements and provides guidance for establishing, implementing, maintaining and continually improving AI management systems within the context of the organization. By understanding trustworthiness as an emergent property of the system, similarly to safety and security, trustworthiness can be improved through an organizational process with specific measurable outcomes and key performance indicators (KPIs); hence, AI trustworthiness can be suitably addressed within the scope of (ISO/IEC 42001:2024). Several initiatives have aimed to propose AI trustworthiness assessment processes, either as an integral part of the safety risk assessment, quality assessment, or data governance processes, or even as a standalone framework of activities. When integrated into a safety assessment process, a risk-based approach is generally adopted to identify, evaluate, and either eliminate or mitigate AI-related hazards and risks. In this context, several studies have focused on identifying and categorizing AI hazards and risks (Sharma, 2024; Zeng et al., 2024), e.g., Zeng et al. (2024) identified 341 AI risks, structured into four categories: System & Operational Risks, Content Safety Risks, Societal Risks, and Legal & Rights Risks.

4. Foundations for AI trustworthy assessment in railway

In this section, we discuss how AI system trustworthiness can be assessed for safe and efficient deployment in railway. Concretely, we aim to adapt and extend the existing railway (safety) risk assessment process in order to take into account the AI trustworthiness (and its attributes). We first recall the safety assessment process and then, we discuss its extension to AI trustworthiness.

4.1. *Railway safety assessment*

Railway systems are sociotechnical systems whose main property is safety. To guarantee and maintain a high level of safety, railway companies (infrastructure managers, railway undertakings, etc.) establish and implement a safety management system. According to the European Union Agency for Railways (ERA), a Safety Management System (SMS) is *the organization, arrangements, and procedures established by a railway company to ensure the safe management of its operations*. The main activities on railway safety management system encompass safety monitoring, investigation, analysis, and reporting of safety occurrences (accidents and incidents), as well as assessing and controlling the associated risks. The ultimate objective of railway risk management is to demonstrate that all identified hazards and risks associated with a proposed change in the railway system are suitably analyzed, evaluated, and reasonably controlled.

Both the European regulation on Common Safety Method for risk assessment[f] (referred to as CSM-RA) and standard (EN 50126:2017)[g] provide a risk management process to be applied to any significant change (impacting safety) on the railway system. Those changes may be technical, operational, and/or organizational, which could impact the operating conditions of the railway system.

The railway risk management process includes (*i*) the risk assessment process, which aims to identify hazards, risks, and associated safety measures and requirements, (*ii*) the demonstration of compliance of the system with the identified

[f] Regulation (EU) No 402/2013 of 30 April 2013 on the common safety method for risk evaluation and assessment.

[g] EN 50126-1 Railway Applications - The Specification and demonstration of reliability, availability, maintainability and safety (RAMS) - Part 1 : generic RAMS process.

safety requirements, and finally, (*iii*) the management of all identified hazards and associated safety measures. The risk assessment process itself includes system definition, risk analysis, and risk evaluation. Risk analysis is derived from the system definition and includes hazard identification, consequence analysis, and selection of the risk acceptance principles. The outcome of such a risk assessment is a set of safety measures and requirements allocated to identified safety functions, systems, or operating rules, aiming to eliminate, mitigate, and control well-defined hazards.

Although initially developed with a primary focus on safety, the railway risk assessment process outlined in standard EN 50126 extends its applicability to other key dependability attributes, including reliability, availability, and maintainability. This is achieved by adapting the definition of risk to encompass concerns related to these properties, thereby enabling a unified approach to assessing and managing risks across all aspects of system dependability. Additionally, the activities related to the dependability management are integrated within the life cycle for the system under consideration, which provides a structure for planning, managing, controlling, and monitoring all the aspects of the system, including dependability.

Similarly, the railway (safety) risk assessment process can be extended to handle AI trustworthiness, by firstly adapting the definitions of hazard/risk to explicitly encompass the various AI-related hazards/risks, integrating trustworthiness within the safety management system, and within the whole life cycle of railway systems.

4.2. *AI trustworthiness assessment framework*

The trustworthiness assessment activities shall be applied for AI railway applications as part of the overall railway system, covering both the development and operational phases. Furthermore, some trustworthiness assessment activities are more related to organizational-level processes, as will be discussed in the sequel. Hence, these activities have to be conducted at different system hierarchical levels (Tonk et al., 2023), namely, (i) *organization level*, (ii) *AI railway application level*, (iii) *AI component level*, and (iv) *AI model level* (see Figure 1).

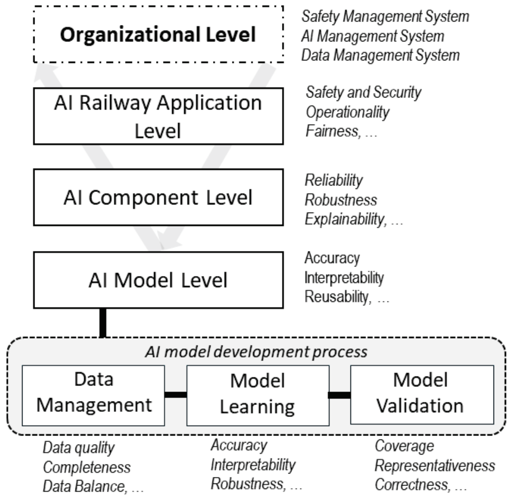

Fig. 1. AI trustworthiness assessment w.r.t. system hierarchical levels.

The *AI railway application level* refers to the AI system deployed in the railway operation environment, e.g., obstacle detection system for autonomous trains, track inspection system, etc. The *AI component level* encompasses a generic AI system, including both its software and hardware components (computation and sensing), e.g., object recognition system, energy optimization system, etc. Finally, the *AI model level* pertains to the AI software algorithm embedded within the AI system, such as a neural network algorithm for object detection. It is essential to consider the AI software within its entire development lifecycle, including data management, training and learning, and validation stages.

As highlighted by Mattioli et al. (2023), trustworthiness extends beyond the intrinsic attributes of the AI product or application; it is in fact deeply associated with the practices, governance, and organizational culture of the entity developing or deploying the system. Therefore, organizations involved in AI development or deployment shall establish an AI system management framework, with trustworthiness and data governance as its core pillars. One of the key activities of trustworthiness management at the organizational level is the continuous evaluation, monitoring, and, when possible, enhancement of the trustworthiness level. This can be achieved by utilizing KPIs to track the levels of high-level attributes of AI trustworthiness.

At the *AI railway application level*, AI trustworthiness analysis should be conducted. This analysis aims to identify AI-specific hazards and risks that may arise during the operation of the railway application within its operational environment. Notice that the considered aspects are intrinsically linked to the risks associated with the system's operating conditions. Consequently, they are often reformulated in terms of "*freedom from*" or "*absence of*" undesirable properties, such as safety, security, ethical concerns, and operability. The standard (VDE-AR-E 2842-61: 2021) provides useful guidelines for such an analysis (Putzer et al., 2021).

Existing catalogs and taxonomies on AI hazards and risks can serve as valuable references to systematically identify those relevant to the system under consideration (Sharma, 2024; Zeng et al., 2024). The identified AI-related risks should then be assessed to determine their acceptability based on predefined risk acceptance criteria. Trustworthiness measures are then established to mitigate and control any risks deemed unacceptable.

The trustworthiness measures, identified at this level, can be linked to high-level trustworthiness attributes. Consequently, they should be considered as goals and requirements that need to be apportioned down to lower system levels. For instance, safety and security, fairness and nondiscrimination, data privacy requirements, etc. Program Confiance.ai (Mattioli et al., 2024) provides an extensive catalog of key trustworthiness attributes that can be leveraged for such a mapping.

While high-level trustworthiness attributes are sociotechnical and more domain-specific, those considered at lower levels are more generic and technical. These lower-level attributes primarily relate to the performances and intrinsic properties of the AI system and its embedded algorithms. Additionally, each high-level attribute can be associated with multiple lower-level attributes, and conversely, a low-level attribute can contribute to several high-level attributes. For instance, safety, as a high-level attribute, may be supported by lower-level attributes such as robustness, explainability, and reliability; and low-level attributes such as explainability can support several high-level ones, such as safety, security, transparency, and ethical alignment.

At the very low level, i.e., AI model level, the analysis should be conducted both quantitatively, using metrics such as accuracy and precision, and qualitatively, as part of broader quality processes. Additionally, it is important to emphasize that this analysis should be proactively integrated into the development process, ensuring that each stage meets specific atomic trustworthiness attributes. For instance, data quality, data completeness, and data balance for the data management stage; accuracy, robustness, reusability, and interpretability for the model learning stage; coverage, representativeness, comprehensiveness for model verification. The work in (Ashmore et al., 2021) provides an initial framework for assessing AI trustworthiness at this lowest level.

Finally, it is important to note that the trustworthiness assessment is an iterative process that combines both top-down and bottom-up approaches. This assessment must be carried out across all system levels to ensure a comprehensive and consistent evaluation.

5. Conclusion & Future works

This paper discusses the trustworthiness of AI systems for deployment in the railway domain and introduces foundations for a framework for their systematic assessment. Concretely, it explores the parallels between railway risk assessment and AI trustworthiness evaluation, refining the definition of risk to account for AI-specific risks and extending the analysis activities to incorporate additional trustworthiness attributes.

While this paper succinctly outlines the high-level process for conducting a trustworthiness assessment, several steps and activities require further elaboration to be fully considered as perspectives. Indeed, at the organizational level, the position and role of the trustworthiness management process should be clarified. For instance, it could be integrated into both AI system management and the safety management system. At a minimum, the interactions between these two management frameworks should be carefully addressed, particularly in the context of AI-based safety-critical applications.

Another important topic, which was beyond the scope of this paper, is the principles and criteria for risk acceptance. As discussed above, the identified AI-related risks should be assessed to determine their acceptability based on predefined risk acceptance criteria. The railway domain has its own (safety) risk acceptance principles, which may be adapted to address AI-related risks.

Finally, the proposed framework needs to be refined and then validated by applying it to key AI-driven railway applications, such as obstacle detection for autonomous trains and track inspection and monitoring.

References

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

Awadid, A., K. Amokrane-Ferka, H. Sohier, J. Mattioli, F. Adjed, M. Gonzalez, and S. Khalfaoui (2024). AI systems trustworthiness assessment: State of the art. In *Workshop on Model-based System Engineering and AI, 12th International Conference on Model-Based Software and Systems Engineering*.

Donato, L. D., F. Flammini, S. Marrone, R. Nardone, and V. Vittorini (2022, June). Trustworthy AI for Safe Autonomy of Smart Railways: Directions and Lessons Learnt from Other Sectors. In *World Congress on Railway Research 2022*, Birmingham, United Kingdom.

European Commission (2020). White paper on artificial intelligence: a european approach to excellence and trust. Accessed: Jan. 25, 2025.

European Commission (2024). Artificial Intelligence Act. Accessed: Jan. 10, 2025.

Jeon, J. (2024). Standardization Trends on Safety and Trustworthiness Technology for Advanced AI. *ArXiv 2410.22151*.

Macrae, C. (2024). Managing risk and resilience in autonomous and intelligent systems: Exploring safety in the development, deployment, and use of artificial intelligence in healthcare. *Risk Analysis*.

Mattioli, J., H. Sohier, A. Delaborde, and al. (2024). An overview of key trustworthiness attributes and KPIs for trusted ML-based systems engineering. *AI and Ethics 4*(1), 15–25.

Mattioli, J., H. Sohier, A. Delaborde, G. Pedroza, K. Amokrane, A. Awadid, Z. Chihani, and S. Khalfaoui (2023). Towards a holistic approach for AI trustworthiness assessment based upon aids for multi-criteria aggregation. In *SafeAI 2023-The AAAI's Workshop on Artificial Intelligence Safety*, Volume 3381.

NIST, A. (2023). Artificial intelligence risk management framework (AI RMF 1.0).

Putzer, H. J., H. Rueß, and J. Koch (2021). Trustworthy AI-based Systems with VDE-AR-E 2842-61. *Proceedings of the Embedded World*.

Sharma, R. (2024). AI Risk Categorization. In *AI and the Boardroom: Insights into Governance, Strategy, and the Responsible Adoption of AI*, pp. 275–286. Springer.

Stettinger, G., P. Weissensteiner, and S. Khastgir (2024). Trustworthiness Assurance Assessment for High-Risk AI-Based Systems. *IEEE Access*.

Tang, R., L. De Donato, N. Besinović, F. Flammini, R. M. Goverde, Z. Lin, R. Liu, T. Tang, V. Vittorini, and Z. Wang (2022). A literature review of artificial intelligence applications in railway systems. *Transportation Research Part C: Emerging Technologies 140*, 103679.

Tonk, A., M. Chelouati, A. Boussif, J. Beugin, and M. El Koursi (2023). A safety assurance methodology for autonomous trains. *Transportation research procedia 72*, 3016–3023.

Wing, J. M. (2021). Trustworthy AI. *Communications of the ACM 64*(10), 64–71.

Zeng, Y., K. Klyman, A. Zhou, Y. Yang, M. Pan, R. Jia, D. Song, P. Liang, and B. Li (2024). AI Risk Categorization Decoded (AIR 2024): From Government Regulations to Corporate Policies. *ArXiv 2406.17864*.