Achieving AI Trustworthiness in Industry – a Policy Perspective

Polina Petrova[1], Denitsa Kozhuharova[1] and Jan Mayer[2]

[1] *Law and Internet Foundation, 54 Bulgarska Morava Str., Fl. 7, Sofia, 1303, Bulgaria*
[2] *Technical University of Berlin, Pascalstraße 8-9, 10587, Berlin, Germany*

Abstract

Exploring the significance and implementation of trustworthiness in artificial intelligence (AI) within the industrial context forms the core theme of this study, highlighting the crucial role of AI's expanding influence in key sectors. It emphasizes key trustworthiness dimensions including robustness, transparency, and fairness, essential for user trust. The work discusses the challenges in implementing these principles, particularly in ethical integration and public trust maintenance. A focus is placed on the EU's AI Act, which introduces a risk-based regulatory framework categorizing AI systems into various risk levels with corresponding regulations. Additionally, the paper examines the relationship between ethics and AI legislation, noting the influence of ethical guidelines on regulatory practices. It then proposes a Trustworthy System Framework for Zero Defect Manufacturing in the industry, incorporating "Trustworthy Pillars", compliance with regulations, robust technical infrastructure, human interaction considerations, process integrity, and operational stability. This framework is designed to enhance the trustworthiness of AI systems in an industrial setting.

Keywords

Trustworthy AI, Industrial AI, AI Policy, Trustworthy Systems, Framework Development

1. What is Trustworthiness Towards AI in an Industrial Context?

The concept of trustworthiness in artificial intelligence (AI) is gaining paramount importance due to the increasing pervasion of AI in industrial applications. This growing importance necessitates the establishment of criteria for validating the quality of AI applications for their intended use, leading to the emergence of the term "trustworthiness" as a set of essential quality requirements for AI applications [1]. The notion of trustworthiness in AI encompasses various dimensions, including robustness, explainability, transparency, reproducibility, fairness, privacy preservation, and accountability. These dimensions play a critical role in building user trust and confidence in AI systems [2].

The operationalization of these principles into practice presents challenges, particularly in integrating ethical considerations into AI development and maintaining public trust. Such integration is indispensable for ensuring sustainable and successful AI innovation [3]. The trust in AI within industrial contexts often relies on two main factors: the perceived reliability of the system and the predictability of its behavior. These factors are of utmost importance for the acceptance of AI in critical sectors such as healthcare, defense, and security [4].

To achieve AI trustworthiness, it is crucial to consider both product and organizational perspectives. This entails conducting AI-specific risk analysis and developing verifiable arguments to establish the trustworthiness of AI applications. Additionally, organizations should implement AI management systems to ensure the responsible handling of AI [1]. Furthermore, the integration of trustworthy AI principles throughout the entire lifecycle of AI systems, from data acquisition to deployment and continuous monitoring, is essential. This comprehensive approach is vital for establishing a foundation of trustworthiness in AI [2].

Proceedings of the Workshop of I-ESA'24, April 10-12, 2024, Chania, Crete (Greece)
EMAIL: polina.petrova@netlaw.bg (P. Petrova); denitsa.kozhuharova@netlaw.bg (D. Kozhuharova); j. mayer@tu-berlin.de (J.Mayer)
ORCID: 0009-0003-6708-2645 (A. 1); 0009-0006-1290-1502 (A. 2); 0000-0002-4734-6768 (A. 3)

2. The AI Act: What Do the Latest Developments Mean for AI-powered Industry?

The rapid progress of AI-driven technologies has led to the urgent need for an adequate and up-to-date legal framework. While aiming to boost research and industrial capacity, ensuring safety and protection of fundamental rights is of supreme importance [5]. Striving to address core grounds and potential risks, EU legislators have thus paved the way to the *Artificial Intelligence Act (*hereinafter "AI Act"*)* [6].

The process of drafting has taken a couple of years, and the EU has worked on different aspects and identified the most vital standards that must be covered by the legislative response. The final steps in this regard were taken in 2023. In June 2023, the European Parliament released its official position, thereby announcing the main requirements to be included: ban of biometric surveillance, emotion recognition, predictive policing, the need of disclosure that content was AI-generated, as well as the high-risk effect of the involvement of AI systems in elections [7].

In December 2023, The Members of the European Parliament (hereinafter "MEPs") reached a political deal with the Council on a Bill to ensure AI in Europe is safe, respects fundamental rights and democracy, while businesses can thrive and expand. The concluded negotiations are a provisional agreement on the AI Act, with the overall unifying goal being to ensure that fundamental rights, democracy, the rule of law and environmental sustainability are protected from high-risk AI, while boosting innovation and making Europe a leader in the field. The rules establish obligations for AI based on its potential risks and level of impact [8].

Over the next few weeks, representatives from the EU institutions shall keep polishing any outstanding technical aspects. In the first half of 2024, the final text shall be presented for approval to the European Parliament and the Council. The approved version shall be published in the Official Journal after it has been translated into the EU's official languages. The implementation period shall then commence twenty days after the publication of the AI Act [9].

It has been agreed that the goal of the EP is to ensure that created and used AI systems are "safe, transparent, traceable, non-discriminatory and environmentally friendly". The prevention of harmful outcomes has also been addressed when it comes to human monitoring. A risk-assessment methodology has been used by the Parliament when approaching this legislation. In other words, the level of potential risks must be addressed and based on that, obligations may be imposed on providers or users. The following paragraphs are dedicated to the types of identified risks [10].

Unacceptable Risk: The AI systems that fall under this category are considered as people threatening and are going to be prohibited. They could include: "Cognitive behavioural manipulation of people or specific vulnerable groups: for example, voice-activated toys that encourage dangerous behaviour in children; Social scoring: classifying people based on behaviour, socio-economic status or personal characteristics; Real-time and remote biometric identification systems, such as facial recognition". Potential exceptions could be introduced such as related to crime prevention [6].

High Risk: The high-risk AI systems are such that can be a treat to safety or fundamental rights. Parties that are creating them should comply with regulations that demand thorough testing, appropriate documentation and a responsibility structure that envisages human oversight. They shall have to be analysed and evaluated prior to being allowed to be distributed, as well as after that. According to the *AI Act*, they shall be divided into 2 main categories [6]:

- **Falling under the scope of EU product safety legislation;**
- **8 areas that are going to have to be registered in an EU database:**
    - **Biometric identification and categorization of people;**
    - **Management and operation of critical infrastructure;**
    - **Education and vocational training;**
    - **Employment, worker management and access to self-employment;**
    - **Access to essential private and public services and benefits;**
    - **Law enforcement;**
    - **Migration, asylum and border control management;**
    - **Legal assistance.**

Limited Risk: Only minimal transparency requirements shall be imposed for this category, with the aim being to give the opportunity to users to make informed decisions. The system must make it clear that users are interacting with AI [6].

Minimal or No Risk: Most of the current AI systems in the EU fall under this category, including applications such as AI-enabled video games or spam filters [6].

Generative AI shall have to implement transparency requirements [6]:

- **AI generated content disclosure;**
- **Designing models to prevent generating illegal content;**
- **Publishing summaries of copyrighted data used for training.**

Apart from assessing the danger posed by AI systems, safeguarding the rights of citizens is also a top priority. Nonetheless, research on AI-based systems may be conducted under open-source licenses to further encourage innovation. Before the system is released to the market, it is integrated into an open test environment for public testing. In addition, the public is urged to make complaints as appropriate and to be aware of choices made using high-risk AI systems that may jeopardize their rights [6].

3. Does Ethics Influence Legislation?

Constituting e set of moral principles, ethics assist in discerning between right and wrong, thereby serving as guidelines for best practice. Engaged with the study of optimizing AI's beneficial impact while reducing risks and adverse outcomes, AI ethics is thus a multidisciplinary field. Ethics has a motivating role during regulatory and legislative processes and oftentimes serves as an inspiration for the creation of new regulatory instruments, for revisions of existing legislation, or for abolishing such. Yet, what ethics is not capable of achieving, but regulation does, is codifying and enforcing ethically desirable behavior [11].

An ongoing debate has been taking place in respect of having ethical guidelines and/or principles navigating AI development instead of actual AI regulation. The two colliding views being that, on the one hand, ethics-driven self-regulation can take the place of external regulation, whilst, on the other, it is recalled that the principles do not provide clarity as to how it may be ascertained that a company follows AI ethics principles. To avoid misunderstanding the role of ethical governance, it must be noted that ethics is not intended to replace regulation. It is a way to conclude what kind of regulation is needed. Therefore, to safeguard that AI ethics possess the ability to uncover ethical issues in a timely manner, thereby serving as a step towards creating appropriate legislation, ethics must be able to accompany the development of AI [11].

With ongoing developments both on national and international fronts, the rather infant legislative framework governing AI is inevitably sure to come. As oftentimes, legislation lags innovation, particularly in the context of AI, proactive measures may still be applied through the help of the principles of ethical use of AI. For example, the *Ethics Guidelines for Trustworthy AI (April 2019)* [12], the *Assessment List for Trustworthy Artificial Intelligence (ALTAI) (end of 2020)* [13], as well as the *White Paper on Artificial Intelligence (February 2020)* [14], have all preceded the *AI Act* [6]. All in all, there is undoubtedly an interplay between ethics and legislation, with ethical considerations underpinning much of the regulatory efforts surrounding AI. In any case, as legislation is constantly evolving, the sound approach to AI ethics calls for always complying with the following principles that form the bedrock for responsible AI development and use:

- **Transparency and Explainability** – AI systems should provide understandable explanations for their decisions and actions, ensuring transparency in their functioning.
- **Fairness & Bias** – to prevent biases and discrimination, AI systems must be developed and trained with a commitment to fairness in decision-making.
- **Accountability** – designers of AI systems should be held accountable for the impact of their creations, fostering responsibility in the development process.
- **Privacy** – respecting user data privacy and handling sensitive information appropriately are fundamental tenets of ethical AI.
- **Safety** – especially in applications like autonomous AI, such as autonomous vehicles and robotics, AI systems should be designed to operate safely, minimizing potential impacts on environments and people.

- **Social Impact** – considerations for positive social impact, including efforts on jobs, economic equality, and social structures, should be integral to AI system design [15].

4. Practical Aspects of Trustworthiness in an Industrial Context

Using specific aspects towards gaining trustworthiness centers its discussion on the establishment of a conceptual framework that is specifically designed to serve as the foundation for ensuring the trustworthiness of systems within the industrial context. This framework is informed by insights gathered from a comprehensive literature review, as well as the analysis of various structural solutions and holistic approaches that are relevant to the concept of Zero-Defect Manufacturing (ZDM) when applied to the industrial setting. The linkage between ZDM and trustworthiness is rooted in the premise that ZDM not only aims to minimize production flaws to the barest minimum but also inherently enhances the reliability and quality of manufacturing processes, thereby fostering trust among stakeholders. Thus, trust is paramount, and it is cultivated through consistent delivery of defect-free products, which signals a company's commitment to excellence and reliability. By integrating advanced technologies, such as real-time monitoring, predictive analytics, and precision engineering, ZDM frameworks provide a proactive approach to identifying and mitigating potential defects before they occur. This proactive behavior not only improves product quality but also significantly reduces the likelihood of costly recalls and reputational damage, further solidifying stakeholder trust.

One of the key components of this framework is the identification and incorporation of what are known as the "Trustworthy Pillars." These pillars are derived from the main characteristics that have been identified through the literature review. Moreover, these pillars align with the different facets of trustworthiness as described by [16]. This standard creates emphasis on the universality of these trustworthiness facets, asserting that they are relevant not only to all systems but also to the core services within the industrial sector as well.

It is important to note that the environment in which a system operates plays a crucial and significant role in determining its overall trustworthiness. This encompasses the system's compliance with various external regulations, including those that are mandated by the EU and other countries where the industrial solutions are deployed. To further elaborate, regulations are defined by the Organisation for Economic Co-operation and Development (OECD) and the Centre for Co-operation with European Economies in Transition [17] as the imposition of rules that are enforced by the government, with penalties being imposed for non-compliance. These regulations, along with the internal and external standards that are in place, are integral components of the Trustworthy System Framework (TSFr) that has been specifically designed for the industrial sector (Figure 1).

The integration of core services into a multi-layered framework, as suggested by [18], efficiently structures software development by segregating functionalities into distinct layers such as advisory, analytical, and data layers. This architecture enhances modularity, facilitating focused development within each layer while maintaining system-wide cohesion. A generic adaptation of this model includes additional layers like edge, storage, and visualization, interconnected through specialized services for data integration, transformation, and secure communication. Key to this framework is the incorporation of security mechanisms and workload optimization solutions, such as blockchain for traceability and AI for edge computing efficiency. This design approach not only simplifies complex software system development but also boosts scalability, flexibility, and security, offering a robust foundation for building dependable software solutions.

Another critical factor which has significant influence on the trustworthiness of a system is its technical infrastructure. The robustness and resilience of the infrastructure against various vulnerabilities, such as attacks, errors, or faults, are vital in instilling trust in the system. To enhance trustworthiness, a multitude of state-of-the-art technologies and methodologies are proposed, which cover a wide range of areas including the development of a comprehensive Data Quality Strategy, the establishment of Data Trustiness and Traceability, the implementation of Data Trusted Communication and Distribution, the assurance of Data Security, and the efficient management of Data Storage and Use.

Furthermore, it is essential to recognize the significance of human factors within the context of semi- automatized industrial solutions. As [19] points out, human interaction with systems poses a potential risk factor that can impact the overall trustworthiness of the system. Therefore, it is of utmost importance to have a thorough understanding of and control over human interaction using conceptual models and methods. Processes, which encompass all hardware and software-related procedures that require human interaction, are identified as crucial elements within the Trustworthy System Framework (TSFr) for the industrial sector. It is important to acknowledge that any flaws or potential faults that may arise within these processes directly impact the overall trustworthiness of the system. Additionally, the use of open-source software libraries in industrial solutions is addressed within the framework. While open-source libraries offer certain benefits, they are often met with skepticism when it comes to trustworthiness. Consequently, a strict quality assurance procedure is mandated for all open-source libraries that are utilized.

The framework also acknowledges the importance of a system's operational integrity in the face of various external and internal events, which range from accidents and attacks to natural disruptions and system failures. In conclusion, the framework incorporates the concept of core services, as proposed by [18], in their Framework for Trusted Software Development. This concept encompasses different layers within the system, such as the advisory layer, analytical layer, and data layer, each of which has its own set of functionalities. These layers collectively form the structure of the entire industrial platform, ensuring interconnectedness and secure communication between various solutions, thereby reinforcing the system's overall trustworthiness.

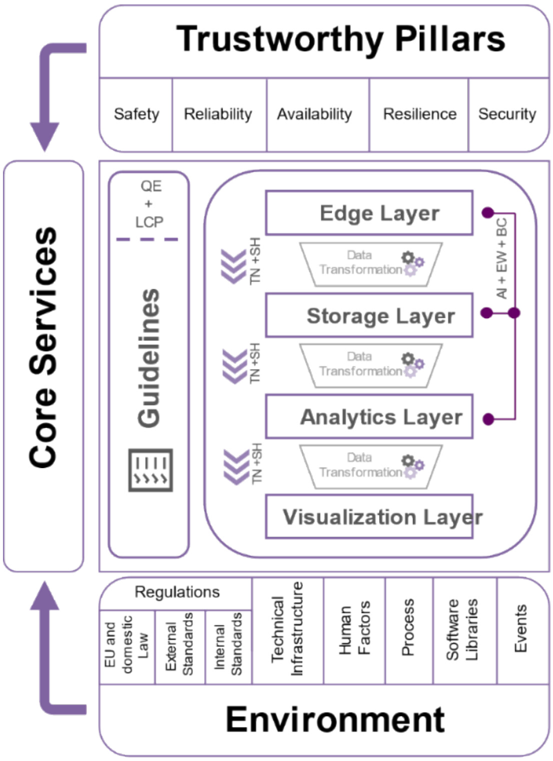

Figure 1: Trustworthy System Framework (TSFr)

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] Schmitz, A., Akila, M., Hecker, D., Poretschkin, M., & Wrobel, S. (2022). The why and how of trustworthy AI. at - Automatisierungstechnik, 70, 793 - 804. https://doi.org/10.1515/auto-2022-0012.

[2] Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2021). Trustworthy AI: From Principles to Practices. *ACM Computing Surveys*, 55, 1 - 46. https://doi.org/10.1145/3555803.

[3] Zhu, L., Xu, X., Lu, Q., Governatori, G., & Whittle, J. (2021). AI and Ethics - Operationalising Responsible AI. *ArXiv*, abs/2105.08867. https://doi.org/10.1007/978-3-030-72188-6_2.

[4] Middleton, S., Letouzé, E., Hossaini, A., & Chapman, A. (2022). Trust, regulation, and human-in-the-loop AI. *Communications of the ACM*, 65, 64 - 68. https://doi.org/10.1145/3511597.

[5] European Commission, "Regulatory framework proposal on artificial intelligence" <https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence> accessed on 29 January 2024.

[6] Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts COM/2021/206 final [2021], available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206>.

[7] European Parliament, "Parliament's negotiating position on the artificial intelligence act" (07.06.2023) <https://www.europarl.europa.eu/thinktank/en/document/EPRS_ATA(2023)747926> accessed on 29 January 2024.

[8] European Parliament, "Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI" (09.12.2023) <https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai > accessed on 29 January 2024.

[9] Council of the European Parliament, "Artificial intelligence act: Council and Parliament strike a deal on the first rules for AI in the world" (09.12.2023) <https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/> accessed on 29 January 2024.

[10] European Parliament, "MEPs ready to negotiate first-ever rules for safe and transparent AI" (14.06.2023) <https://www.europarl.europa.eu/pdfs/news/expert/2023/6/press_release/20230609IPR96212/20230609IPR96212_en.pdf> accessed on 29 January 2024.

[11] European Parliament. "Artificial intelligence: From ethics to policy", Study Pantel for the Future of Science and Technology (June, 2020) < https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641507/EPRS_STU(2020)641507_EN.pdf > accessed on 18 April 2024.

[12] Independent High-Level Expert Group on Artificial Intelligence Set up by the European Commission, "Ethics Guidelines for Trustworthy AI. High-Level Expert Group on Artificial Intelligence" [2019], available at <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>.

[13] Independent High-Level Expert Group on Artificial Intelligence Set up by the European Commission, "Assessment List for Trustworthy Artificial Intelligence (ALTAI)" [2020], available at: < https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment >.

[14] European Commission, "White Paper on Artificial Intelligence – A European approach to excellence and trust" [2020], available at: < https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en >.

[15] Thilo Hagendorff, "A Virtue-Based Framework to Support Putting AI Ethics into Practice" (21 June 2022) , available at: <https://link.springer.com/article/10.1007/s13347-022-00553-z>.

[16] British Standards Institution (2018). Information technology - Systems trustworthiness - Part 1: Governance and management specification (BS 10754-1:2018).

[17] OECD; Centre for Co-operation with European Economies in Transition. (1993). Glossary of industrial organisation economics and competition law. OECD.

[18] Bose, R. P. J. C., Singi, K., Kaulgud, V., Phokela, K. K., & Podder, S. (2019). Framework for Trustworthy Software Development. In 2019 34th IEEE/ACM International Conference on Automated Software Engineering Workshop (ASEW) (pp. 45–48). IEEE. https://doi.org/10.1109/ASEW.2019.00027.

[19] Schneider, F. B. (Ed.). (1999). Trust in Cyberspace. The National Academies Press. https://doi.org/10.17226/6161.