# AI Certification: Empirical Investigations into Possible Cul-De-Sacs and Ways Forward

**Benjamin Fresz**
Fraunhofer Institute for Manufacturing Engineering and Automation IPA, Stuttgart, Germany
Institute of Industrial Manufacturing and Management IFF, University of Stuttgart, Germany

**Danilo Brajovic**
Fraunhofer Institute for Manufacturing Engineering and Automation IPA, Stuttgart, Germany
Institute of Industrial Manufacturing and Management IFF, University of Stuttgart, Germany

**Marco F. Huber**
Fraunhofer Institute for Manufacturing Engineering and Automation IPA, Stuttgart, Germany
Institute of Industrial Manufacturing and Management IFF, University of Stuttgart, Germany

## Abstract

In this paper, previously conducted studies regarding the development and certification of safe Artificial Intelligence (AI) systems from the practitioner's viewpoint are summarized. Overall, both studies point towards a common theme: AI certification will mainly rely on the analysis of the processes used to create AI systems. While additional techniques such as methods from the field of eXplainable AI (XAI) and formal verification methods seem to hold a lot of promise, they can assist in creating safe AI-systems, but do not provide comprehensive solutions to the existing problems in regard to AI certification.

**2012 ACM Subject Classification** Social and professional topics → Testing, certification and licensing; Computing methodologies → Machine learning; General and reference → Empirical studies

**Keywords and phrases** AI certification, eXplainable AI (XAI), safe AI, trustworthy AI, AI documentation

**Digital Object Identifier** 10.4230/OASIcs.SAIA.2024.13

**Category** Practitioner Track

**Funding** This paper is funded in parts by the German Federal Ministry for Economic Affairs and Climate Action under grant no. 19A21040B (project "veoPipe"), an OEM company via a joint project and by the Fraunhofer Gesellschaft under grant no. PREPARE 40-02702 (project "ML4Safety").

**Acknowledgements** Parts of this paper were refined with the help of company-specific LLM (Fh-Genie 4o).

## 1 Introduction

Artificial Intelligence (AI) has rapidly integrated into various industries, necessitating the development of robust certification standards to ensure reliability, safety, and ethical compliance. Current challenges include the lack of standardized guidelines and the opaqueness of AI decision-making processes, which can lead to mistrust and potential misuse [4, 5, 7]. AI certification could – when done correctly – ensure the performance and trustworthiness of AI systems. Concurrently, Explainable AI (XAI) addresses the need for transparency and interpretability in AI models, aiming to make AI decisions comprehensible to humans, foster trust, and enable informed decision-making.


## 2 The EU AI Act

The European Union's AI Act [1] proposes a regulatory framework to govern AI systems, emphasizing transparency, accountability, and risk management. The AI Act categorizes applications into different risk levels, each requiring specific compliance measures. However, the AI Act lacks technical details, which are expected to be specified in future harmonized standards. This regulatory uncertainty underscores the need for practical insights and frameworks to guide the effective certification of AI systems in real-world settings. For example, such guidelines could provide clear information on how to document the development process of AI applications or the selection process of appropriate training data. Industry practitioners and researchers can contribute valuable insights by sharing experiences and developing best practices that align with the goals of transparency, accountability, and risk management.

## 3 Empirical Studies

Two studies conducted in Germany aimed to gather such insights [2, 3]. The first study evaluated general opinions regarding AI safeguarding within industry, while the second focused on the use of XAI methods for certification purposes.

### 3.1 Effectively Documenting AI Applications

The first study reviewed activities surrounding the safeguarding and certification of AI systems over recent years. Collaborations with industry partners in Germany across various domains (automotive, finance, and food manufacturing) and interviews with certification experts, employees, and developers led to the publication of a framework for documenting AI applications along four major development steps [2]. The feedback from practitioners highlighted five main points:

*The AI Assessment Catalogue* [6] is widely used in Germany and regarded as a de facto standard, but its length (approximately 160 pages) poses practical challenges. It is primarily designed for final assessments, not for providing support during the development phase.

There is some *uncertainty about the Assessment Catalog,* as practitioners are concerned about its compliance with the AI Act and its extensive length, though it remains the primary tool for developing safe AI.

Partners expressed *high hopes for future standards,* as they are eagerly awaiting new standards for safe AI development, showing interest in best practices for data collection and model selection, despite there being some concern about the practical usefulness of these standards.

*Need for implementation details, technical tools, and clearly defined performance thresholds* was often expressed, as partners desire more concrete guidance, including specific tools and methods, actionable instructions, and specific information to facilitate the development process.

*Trust-Building is a key motivation* to pursue certification of AI systems. Affected individuals want to be included in the development process and understand the design choices made.

## 3.2 XAI for Safe AI and Certification

The second study focused on experts in XAI and AI certification to explore expectations regarding the use of XAI in this area [3]. Through 15 qualitative interviews, it was found that XAI can help to identify errors such as biases within AI models. However, it cannot comprehensively answer the question, "Is this AI model safe to use?"

Surrounding the use of XAI, some common themes (also in part echoing the practitioners' view of the first study) emerged: Experts missed clear guidance and standardization on how and when to use XAI. They employed a variety of XAI methods to address specific problems, sometimes successfully (especially as a communication tool between experts) and sometimes unsuccessfully (when data relationships were unknown or too complex). Additionally, XAI methods face several challenges:

- Difficulty in implementation or use.
- Growing complexity of AI systems, such as large language models (LLMs).
- Need to adapt XAI methods to different data types, use cases and AI models, including multi-modal ones.
- Difficulty (or impossibility) in objectively assessing the quality of explanations.

These challenges overall led the experts to believe that XAI is unlikely to provide comprehensive solutions to AI certification in the near future. While some hope was levied towards new XAI-approaches, the most promising methods mentioned regarding AI certification (apart from just spotting ML model biases) were formal verification methods that provide (statistical) guarantees for AI properties and AI approaches incorporating transparent decision-making by design, such as neuro-symbolic approaches.

## 4 Summary

AI certification and Explainable AI are crucial for ensuring the reliability, safety, and ethical compliance of AI systems. The European Union's AI Act aims to regulate AI systems by emphasizing transparency, accountability, and risk management, though technical details are still needed. Industry practitioners and researchers can contribute valuable insights for practical frameworks and best practices. Two German studies highlight the need for detailed guidance, robust standards, and trust-building in AI certification. While XAI methods are sometimes mentioned as a possible solution to AI certification, they cannot fully guarantee AI safety but can be valuable tools for identifying biases. As such, they can be used as additional assets in development and certification processes.

## References

1 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance), 2024. URL: http://data.europa.eu/eli/reg/2024/1689/oj.

2 Danilo Brajovic, Niclas Renner, Vincent Philipp Goebels, Philipp Wagner, Benjamin Fresz, Martin Biller, Mara Klaeb, Janika Kutz, Jens Neuhuettler, and Marco F. Huber. Model Reporting for Certifiable AI: A Proposal from Merging EU Regulation into AI Development, 2023. arXiv:2307.11525, doi:10.48550/arXiv.2307.11525.

3 Benjamin Fresz, Vincent Philipp Göbels, Safa Omri, Danilo Brajovic, Andreas Aichele, Janika Kutz, Jens Neuhüttler, and Marco F. Huber. The Contribution of XAI for the Safe Development and Certification of AI: An Expert-Based Analysis, 2024. arXiv:2408.02379, doi:10.48550/arXiv.2408.02379.

4 Yueqi Li and Sanjay Goel. Making It Possible for the Auditing of AI: A Systematic Review of AI Audits and AI Auditability. *Information Systems Frontiers*, 2024. doi:10.1007/s10796-024-10508-8.

5 Jakob Mökander. Auditing of AI: Legal, Ethical and Technical Approaches. *Digital Society*, 2(3), 2023. doi:10.1007/s44206-023-00074-y.

6 Maximilian Poretschkin, Anna Schmitz, Maram Akila, Linara Adilova, Daniel Becker, Armin B. Cremers, Dirk Hecker, Sebastian Houben, Michael Mock, Julia Rosenzweig, Joachim Sicking, Elena Schulz, Angelika Voss, and Stefan Wrobel. Guideline for Trustworthy Artificial Intelligence – AI Assessment Catalog, 2023. arXiv:2307.03681, doi:10.48550/arXiv.2307.03681.

7 Joyce Zhou and Thorsten Joachims. How to explain and justify almost any decision: Potential pitfalls for accountability in ai decision-making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 12–21, New York, NY, USA, 2023. Association for Computing Machinery. doi:10.1145/3593013.3593972.