



Autischer Gregor

Practical Application and Limitations of AI Certification Catalogues

Bachelor-Thesis to achieve the university degree of Bachelor of Science
Bachelors degree programme: Information and Computer Engineering

submitted to

Graz University of Technology

Supervisor Dipl.-Ing. Dr.techn. BSc Dominik Kowald
Dr. Kerstin Waxnegger

Institute of Interactive Systems and Data Science
Head: Univ.-Prof. Dipl.-Ing. Dr.techn. Frank Kappe

Graz, October 2024

Abstract

This work investigates the certification of artificial intelligence systems (AI systems), focusing on the practical application and limitations of existing certification catalogues by attempting to certify a publicly available AI system. The study aims to evaluate how well current approaches, such as the Fraunhofer AI Assessment Catalogue, work to effectively certify an AI system. And how publicly accessible AI systems, that might not be actively maintained or initially intended for certification, can be selected and used for a sample certification. The methodology involves implementing the Fraunhofer AI Assessment Catalogue as a comprehensive tool to systematically assess an AI model's compliance with certification standards. Results showed that while the catalogue effectively structures the evaluation process, it can also be cumbersome and time-consuming. It also showed the importance of complete system documentation, which is particularly important, if the system is not actively developed. The limitation of not having a development team available for system-related questions and documentation improvements became evident. Some limitations of the used certification catalogues also became apparent, and thoughts as to how to streamline a certification process are presented. This work practically demonstrates how a certification could be done and what challenges arise on the way. It particularly focuses on how an AI system should be selected for a sample certification and the shortcomings such an approach might have.

Contents

1	Introduction	1
2	Current State of Regulation	3
2.1	EU Artificial Intelligence Act	3
2.1.1	Scope	4
2.1.2	Definition of AI	4
2.1.3	AI risk categories	5
2.2	EU Artificial Intelligence Liability Directive	6
2.3	Other Global Initiatives	6
3	Certifying AI	7
3.1	Certification Catalogues	8
3.1.1	Fraunhofer AI Assessment Catalogue	8
3.1.2	Trusted Artificial Intelligence: Towards Certification of Machine Learning Applications	9
3.1.3	Auditing Machine Learning Algorithms	10
4	Facial Emotion Recognition - AI Application	11
4.1	Decision to Use EmoPy Framework and RIOT Project	11
4.2	Brief Overview of the EmoPy Framework	13
4.3	Brief Overview of the RIOT Project	13
4.4	The complete AI application	15
5	Approach	16
5.1	Before the certification process	17
5.2	AI Profile	18
5.3	Life Cycle of the AI Application	18
5.4	Protection Requirement Analysis	18
5.5	Risk Analysis	19
6	Findings	20
6.1	Selecting the AI-System	20
6.1.1	Initial Considerations and Challenges	21

Contents

6.1.2	Context and Embedding Requirements	21
6.1.3	Documentation Requirements	21
6.2	The Certification Process	22
6.2.1	Challenges During Certification	22
6.2.2	Specific Observations on the Fraunhofer Catalog	23
6.3	Learnings	24
6.3.1	Catalogue-Specific Observations	24
6.3.2	Limitations	24
6.3.3	Recommendations for future Sample Certifications	25
6.3.4	Recommendations for Real-World Applications	25
7	Conclusion	26
	The Facial Emotion Recognition System	32
1	RIOT Installation	32
2	EmoPy Framework	32
3	The AI Application that is certified	33
	AI Profile	34
	AI life cycle	38
	Protection Requirement Analysis	40
1	Dimension: Fairness	40
2	Dimension: Autonomy and Control	41
3	Dimension: Transparency	42
4	Dimension: Reliability	43
5	Dimension: Safety and Security	44
6	Dimension: Data Protection	45
	Risk Analysis	47
1	Dimension: Reliability	47
1.1	Reliability in standard cases	47
1.2	Robustness	49
1.3	Interception errors at model level	52
1.4	Uncertainty estimation	54
1.5	Control of dynamics	55
1.6	Summary of the dimension	57
2	Dimension: Fairness	57
2.1	Risk area: Fairness	57
2.2	Risk area: Control of dynamics	60

Contents

2.3	Summary of the dimension	61
3	Cross-dimensional assessment	61
4	Conclusion	62

List of Figures

4.1	RIOT Installation in New York 2018	14
-----	--	----

1 Introduction

Artificial intelligence (AI) has a long history, dating back to the mid-20th century when the term was first adopted. Over several decades, AI has evolved to complex machine learning (ML) algorithms and neural networks that are common place today. However, it is in recent years that AI systems have become increasingly prevalent in our daily lives, moving beyond specialized research labs and into mainstream applications (Costa and Aparicio, 2023). From virtual assistants on our smartphones to recommendation systems in e-commerce, AI now plays a significant role in shaping our interactions with technology and informing decision-making processes across various sectors (Kasinidou et al., 2024). The rapid proliferation of AI applications has raised concerns about safety, privacy, fairness, and ethical implications (Baum et al., 2023).

In response to these challenges, AI governance has become an increasingly prominent focus for legislators and policymakers worldwide (Smuha, 2021). The European Union’s AI Act, for instance, represents a landmark piece of legislation that aims to establish a comprehensive regulatory framework for AI systems (Mueck et al., 2022). Similar initiatives are underway in other countries and regions, reflecting a growing global consensus on the need for AI governance (Nad et al., 2023). The necessity for AI system certification is rising in importance to achieve regulatory compliance and foster confidence in artificial intelligence technologies. Given that the EU AI Act took effect on August 1, 2024, this has now become more important than ever.

The certification of AI systems represents a complex and relatively new challenge, distinctly different from traditional software certification. While conventional software certification primarily focuses on functionality and security, AI certification must address a broader spectrum of concerns, including performance accuracy, fairness, transparency, and ethical considerations (Winter et al., 2021).

This work aims to bridge the gap between theoretical certification frameworks and their practical application, by attempting to certify an existing open-source AI system using current frameworks and documenting the successes, failures, and challenges encountered along the way. Fraunhofer’s AI Assessment Catalogue by Poretschkin et al. (2023) will be used to guide through the certification process. Comparisons to other catalogues are also made. This approach should lead to a better understanding of:

1 Introduction

- Identify which parts of the catalogue are most useful and if simplifications or refinements could be beneficial
- Provide a more practical understanding of the potential certification process
- Discover the limitations where sample certifications encounter challenges

This work provides a comprehensive walk through of an AI certification in practice. It starts with essential background information on the current landscape of AI regulations and existing certification catalogues. It then details the specific AI system chosen for certification, including the preparations that were needed to be made. The core of the thesis documents the certification attempt, highlighting both successes and challenges. It offers an analysis of what aspects were achievable and which proved problematic. The thesis concludes by describing these findings and offering recommendations for future developments in AI certification.

2 Current State of Regulation

The rapid advancement and widespread adoption of AI across various industries have necessitated the development of comprehensive regulatory frameworks. While early applications like automatic image sorting in photo libraries posed minimal risk, the increasing deployment of AI in more critical areas necessitates stricter regulations (Pimentel, 2024). Some AI applications, particularly in medicine, have already received approvals from bodies like the US Food and Drug Administration (Benjamens et al., 2020). In the European Union some medical applications also got approval, one example is the ChestLink software, that automatically reports normal chest X-Rays. Systems such as this set an important precedent for other automated medical and potentially nonmedical systems in the future (Saenz et al., 2023). Regulating AI-powered systems is not a recent development. However, contemporary approaches are primarily concerned with specific domains and lack general applicability, they could provide a blueprint for other applications. Recently, lawmakers have recognized the urgent need for comprehensive regulations. These frameworks intend to encompass not only individual industries, but also to create universal rules that ensure consistency and fairness. Worldwide, several policy initiatives are in progress (Nad et al., 2023). Although there are many more not mentioned here, the following are some key efforts in this area.

2.1 EU Artificial Intelligence Act

The European Union's AI Act is currently the most significant and far-reaching regulatory initiative in the field of AI. Its impact is expected to extend beyond the EU's borders, potentially setting global standards for AI management (Mueck et al., 2022). It was recently finalized and has been in effect since August 1, 2024. Key Objectives of the AI Act are (European-Union, 2024):

- **Ensure AI Safety and Compliance:** Guarantee that AI systems in the EU are safe and adhere to laws protecting fundamental rights and EU values, safeguarding users' privacy and preventing discrimination.

2 Current State of Regulation

- **Certainty for AI Investment:** Establish a clear legal framework to foster innovation and investment in AI, providing businesses and investors with regulatory clarity.
- **Enhance Governance:** Strengthen enforcement mechanisms to effectively apply existing laws on fundamental rights and AI safety requirements.
- **Unify AI Market:** Foster a single, cohesive market for trustworthy AI applications across the EU, preventing fragmentation through harmonized regulations.

The legislation uses a risk-based approach to achieve these goals.

2.1.1 Scope

The scope of the AI Act, as detailed in Article 2 of European-Union (2024), is broad and inclusive, covering various actors and scenarios within the AI ecosystem. The regulation applies to providers, deployers, importers, and distributors of AI systems, as well as product manufacturers incorporating AI systems into their products, regardless of their location, if the AI system or its output is used within the European Union. It encompasses high-risk AI systems, general-purpose AI models, and systems that may impact fundamental rights. The AI Act explicitly excludes AI systems developed purely for scientific research and development. It also excludes AI systems published under open-source licences. However, regulation applies if an open-source system is brought to market or is in use as a high-risk AI system. It respects existing EU laws on data protection and consumer safety, while allowing for stricter national regulations in areas like worker protection. The AI Act imposes obligations across the entire AI value chain, ensuring a comprehensive approach to AI governance and safety within the European market.

2.1.2 Definition of AI

The AI Act adopts a broad and technology-neutral definition of AI systems, focusing on their functional characteristics rather than specific technologies or methods. According to the AI Act, an AI system is defined as a machine-based system designed to operate with varying levels of autonomy. It potentially exhibits adaptiveness after deployment, and is capable of generating outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments based on input it receives (European-Union, 2024). This definition emphasizes two key elements: 'inference' and 'autonomy', which distinguish AI systems from traditional software with predetermined outputs. The AI

Act takes a broad approach to stay relevant as technology rapidly advances. It encompasses both the core AI system and its surrounding code, recognizing the complexity of AI applications. This definition aligns with the OECD's latest conceptualization of AI and moves away from earlier, more restrictive definitions tied to specific technologies (Fernhout and Duquin, 2024). By focusing on the system's ability to operate autonomously, adapt, and influence environments based on inferred objectives, the AI Act creates a framework that can accommodate current and future AI technologies, ensuring its long-term applicability in regulating the AI landscape.

2.1.3 AI risk categories

The AI Act categorizes AI systems into various risk levels and imposes corresponding requirements, with stricter regulations for higher-risk applications (European-Union, 2024). Key categories include:

1. Prohibited AI Practices
2. High-Risk AI Systems
3. General-Purpose AI Models
4. Transparency-Obligated Systems
5. Limited-Risk Systems

Prohibited practices include, but are not limited to, systems that manipulate behaviour, exploit vulnerabilities, or create facial recognition databases from untargeted scraping. High-risk AI systems are defined as such, if they pose significant risks to health, safety, or fundamental rights, and are either used as products (or components of products) covered by specific EU legislation or listed in Annex III of the AI Act. Annex III includes several areas such as biometric identification, emotion recognition systems, management of critical infrastructure, education, employment, and law enforcement. General-purpose AI models face some regulation, but it is less stringent than for high-risk systems. The AI Act also provides an exception for certain AI systems that, despite falling under Annex III categories, may not be considered high-risk. This is the case if they perform narrow procedural tasks or do not significantly influence human decision-making, provided they do not involve profiling of natural people. Transparency-Obligated Systems, according to Article 50, require clear disclosure when users interact with AI or encounter AI-generated content. Limited-Risk Systems, according to Article 95, are applications with minimal restrictions, that are encouraged to follow voluntary codes of conduct for ethical and responsible use.

2.2 EU Artificial Intelligence Liability Directive

Complementing the AI Act, the European Commission proposed the AI Liability Directive in September 2022. This directive aims to modernize and enhance the EU's liability framework for AI systems (Madiaga, 2023). The directive aims to ensure that individuals who suffer damages from AI systems receive equivalent protection to those harmed by other forms of technology. It incorporates a rebuttable presumption of causality, which simplifies the process for victims by presuming a causal link between an AI system's fault and the damage caused. Additionally, it empowers national courts to mandate the disclosure of evidence concerning high-risk AI systems, aiding victims in substantiating their claims. By standardizing liability regulations across the EU, the directive aims to avoid legal discrepancies and guarantee uniform protection for those impacted by AI-related damages. Coordinating with the AI Act and other EU measures, the AI Liability Directive addresses only non-contractual liability claims, encompassing a wide array of potential AI-related harms (European-Union, 2022). This comprehensive strategy seeks to balance the protection of victims with the encouragement of AI innovation, reducing legal uncertainties and promoting the responsible advancement of AI technologies within the EU.

2.3 Other Global Initiatives

While the EU's regulatory efforts are the most comprehensive, other countries and regions have also proposed AI regulations. The following examples represent a small sample of international regulatory efforts and are not an exhaustive list. The US has introduced a Blueprint for an AI Bill of Rights, which outlines principles for the design and deployment of AI systems (White-House, 2022). The UK has proposed a sector-led approach to AI regulation. Japan has developed AI Guidelines emphasizing a multi-layered governance framework (Department for Digital, 2022).

Unlike the legally binding EU AI Act, these initiatives do not have legal enforceability. They still highlight the global acknowledgment of the necessity for AI regulation and certification standards. The current state of AI regulation and certification is rapidly evolving, with the EU taking a leading role through its comprehensive AI Act and AI Liability Directive. These regulations are poised to shape global standards for AI development, deployment, and governance in the coming years.

3 Certifying AI

Certification is an important and established tool to prove that technical systems meet certain standards or regulations. Successful certifications are not only vital to prove compliance of a system to according norms. But it is also crucial in helping to establish trust in a system from its users (Blösser and Weihrauch, 2023). The process of certifying systems has a long history and differs slightly from application to application. For traditional software projects where every line of code can be reviewed one by one, certification processes have been established (Winter et al., 2021).

However, the certification of AI applications is currently in its early stages. This is partly because comprehensive legal frameworks for AI have only recently begun to be developed. A prime example is the AI Act, which officially came into effect on August 1, 2024. Despite ongoing efforts by international standardization organizations like ISO, IEC, and IEEE to create guidelines and standards, a fully established certification process has yet to be achieved (Nad et al., 2023). These efforts aim to address the distinct challenges presented by AI technologies. The establishment of robust certification processes for AI has been hindered by the absence of comprehensive legal frameworks. Even with recent legislative developments, the rapidly evolving nature of AI technology continues to pose significant challenges for creating and maintaining effective certification standards (Delgado-Aguilera Jurado et al., 2024). This dynamic landscape requires certification processes that are both adaptable and rigorous, capable of evolving alongside the technology they aim to regulate.

Certifying AI systems also presents unique challenges due to their complex and often opaque nature. Many decision-making processes are not directly programmed by humans. Instead, these systems use various methods to analyse and interpret data, developing their own decision-making processes. Humans do have multiple ways to influence and understand how the system makes decisions and how good the outcome is. But the process is not completely transparent. This complexity necessitates a different approach to certification, emphasizing the need to establish a comprehensive framework that accounts for the inherent opacity and adaptability of AI technologies (Winter et al., 2021).

Moreover, AI systems are capable of evolving and changing their behaviour over time as they can be retrained with new data. This continuous learning process adds

another layer of complexity to certification, as it requires ongoing monitoring and re-evaluation to ensure that the systems remain compliant with ethical standards and avoid biases. Addressing these dynamic aspects is essential to develop robust certification practices that can keep pace with the evolving nature of AI (Benjamin Fresz et al., 2024).

3.1 Certification Catalogues

CEN, CENELEC and ETSI are leading European standardization bodies, they bridge the gap between EU regulations and practical certification frameworks designed to evaluate and certify AI systems. They integrate these guidelines with European legislative priorities, and ensure consistency across the European standardization landscape (Hadrien Pouget, 2024). Several organizations have created catalogues and guidelines to evaluate, test, and certify AI systems. Prominent examples include the Fraunhofer AI Assessment Catalogue by Poretschkin et al. (2023), the white paper “Trusted Artificial Intelligence” from TÜV Austria and Johannes Kepler University Linz by Winter et al. (2021), and the white paper “Auditing Machine Learning Algorithms” by the Supreme Audit Institutions of various countries (SAI-FI-DE-NL-NO-UK, 2023). These frameworks provide distinct methodologies and list criteria to certify AI applications, addressing aspects such as fairness, autonomy and control, transparency, reliability, safety and security, and data protection. This work has its focus primarily on the Fraunhofer Certification-Catalogue and how it applies to an AI system.

3.1.1 Fraunhofer AI Assessment Catalogue

The Fraunhofer AI Assessment Catalogue describes the important role of quality assurance in artificial intelligence (AI) as they become increasingly integrated into various sectors of society. The paper emphasizes the necessity of implementing stringent quality standards to ensure AI systems are reliable, safe, aligned with societal values and compliant with the law, particularly in sensitive application contexts (Poretschkin et al., 2023). The paper identifies several key challenges in assessing and ensuring AI quality. These include the complex value chain involved in AI development, the intricate nature of AI systems, and the difficulty in comprehending the inner workings of AI models. The paper argues that these challenges necessitate a systematic approach to quality implementation in AI development and highlights the importance of unbiased expert assessment in establishing trust in AI applications. A significant focus of the catalogue is on the operationalization of quality requirements for AI. The paper recognizes that although there are

established guidelines for reliable AI, the specifics of how they can be practically applied remain mostly unclear. It references the European Commission’s draft regulation on AI, which mandates conformity assessments for high-risk AI systems, and discusses the need for specific, quantifiable criteria for assessing AI quality. The paper proposes a risk-based AI assessment approach and introduces an AI assessment catalogue. This catalogue provides a structured approach for certifying AI applications across six dimensions of trustworthiness: fairness, autonomy and control, transparency, reliability, safety and security, and data protection. The proposed framework aims to complement existing assessment approaches by offering a more concrete procedure for developing safeguarding arguments for AI applications. It tries to support developers and operators of AI systems in meeting regulatory requirements. The information it provides can be used create technical documentation for a conformity assessments. It introduces a Step-by-step process that this work uses when trying to certify an AI System.

3.1.2 Trusted Artificial Intelligence: Towards Certification of Machine Learning Applications

This White paper published by TÜV Austria and Johannes Kepler University Linz wants to outline a structured approach to certifying machine learning applications (Winter et al., 2021). It describes key ML principles and discusses relevant aspects and challenges in the context of certification. It emphasizes that while ML systems are complex, they are not black boxes, but rather white boxes whose operations can be analysed in detail. The authors highlight the risks associated with the improper use of ML techniques, especially given the proliferation of user-friendly ML tools. They also stress the importance of proper scientific methodologies when creating and operating such systems. The paper introduces a certification approach for ML applications, focusing initially on supervised learning tasks with low-risk potential. Part of an audit catalogue is presented, as a foundation for this certification process. It is designed to be refined and expanded based on practical experience and scientific developments. The key aspects this paper focuses on are technical aspects, such as evaluating the technical robustness and reliability. It also focuses on ethical considerations and cybersecurity measures that need to be taken. The paper presents a summary of the certification framework created by TÜV Austria, intended to offer comprehensive instructions for evaluating AI systems across various aspects. However, the actual catalogue is not publicly accessible, which means it can’t be directly used to certify an AI model. However, it can serve as a reference point, that highlights key areas to consider during the certification process.

3.1.3 Auditing Machine Learning Algorithms

A collaboration of European public auditing institutions released this White paper. It describes the increasing use of AI/ML in public services, which necessitates new certification methodologies (SAI-FI-DE-NL-NO-UK, 2023). It identifies several risks, including an over-focus on numeric metrics at the expense of compliance and fairness, miscommunication between product owners and developers, over-reliance on external expertise, and uncertainty regarding personal data use.

To address these challenges, the paper proposes a certification framework covering the entire AI application lifecycle. The audit areas focus on data understanding, model development, performance, and ethical considerations such as explainability and fairness. The paper emphasizes the need for auditors to develop specific competencies in ML principles, coding languages, and cloud-based infrastructures. To aid in this process, the paper introduces a helper tool, in the form of an Excel Sheet. It is meant for auditors to prepare and conduct AI audits efficiently. The authors stress that specialized knowledge and skills are required for ML certifications. They emphasize that the proposed audit catalogue and helper tool should be continuously refined and updated. Ultimately, this paper aims to provide guidance and good practices to enable auditors to navigate different parts of the certification process, helping them prepare and execute certifications effectively.

4 Facial Emotion Recognition - AI Application

To undertake a certification process, the first requirement is a system that either requires certification or is eligible for it. The system should incorporate an AI component, ideally leveraging machine learning techniques. Furthermore, it is essential that the AI component is integrated within a larger, comprehensive system, as certifications are typically applicable only to entire systems rather than isolated components. Specifically, the Fraunhofer Certification Catalogue mandates a well-defined assessment object, which can only be established when the AI component is part of a larger, integrated system (Poretschkin et al., 2023).

Addressing these challenges involves identifying and selecting a suitable system that meets these criteria. This system must not only include a machine learning-based AI component but also demonstrate a level of complexity and integration that warrants a certification. The objective is to find a system where the AI component plays a critical role in the overall functionality, thereby making the certification process relevant and meaningful.

In essence, the aim is to identify a system that incorporates a machine learning AI module within a broader architecture.

4.1 Decision to Use EmoPy Framework and RIOT Project

After careful consideration, the decision was made to use the EmoPy Framework (Angelica Perez et al., 2021) and its implementation within the RIOT Project (Thoughtworks, 2018a) for this certification study. Several key factors influenced this choice, ensuring that both the framework and the project are well suited for the certification process.

Firstly, the relevance to the EU AI Act played a significant role in the decision-making process. Emotion recognition, the primary focus of the EmoPy Framework, is potentially covered under the EU AI Act. Although open-source models like EmoPy are generally not covered by this regulation, such a system could be covered if it was brought to market.

Another critical factor was the open-source nature and transparency of EmoPy. EmoPy is fully open-source, which enables a in-depth look into the technical details and makes the system fully transparent, which should make the certification process possible. The codebase of EmoPy is relatively small and manageable, making it easier to understand and verify. Yet, it is sufficiently large to make a certification worthwhile. Additionally, the framework is well-documented, with multiple articles and resources that describe the model selection process. This level of documentation is critical for certification, as it provides clear insights into the design and functionality of the model, facilitating a thorough evaluation. While solid documentation is essential for any certification, in this particular case study, it is especially critical because there is not an active company or team of developers working on the project, as one would expect in a real-world scenario. Although there is not an active development team for ongoing support, contacting Angelica Perez, the author of the EmoPy framework articles and its lead developer, proved beneficial as she kindly addressed questions about the framework and the RIOT setup, somewhat mitigating this drawback. Other than that, all necessary information for certification must be extracted from the provided documentation. This presents a limitation compared to a standard certification process, where ongoing interactions with active developers are commonplace.

From a technical perspective, the suitability of the EmoPy Framework for a sample certification was also a major consideration. The framework is well-suited to the certification process due to its technical characteristics and the comprehensiveness of its documentation. Moreover, the integration of EmoPy within the RIOT Project provides a complete system context, which is essential for certification. Validating standalone machine learning models can be difficult with traditional certification catalogues. But the RIOT Project provides a comprehensive setting where the AI component is integrated into a broader system. This integration is crucial, as a certification typically applies to a entire systems rather than individual components.

The decision to use the EmoPy Framework and the RIOT Project for this certification study was driven by all these factors. The potential relevance to the EU AI Act, it being open-source and therefore transparent, and the suitability of the framework for a sample certification all contributed to make this choice. The comprehensive system context provided by the RIOT Project further strengthened the decision, ensuring that the certification process can be conducted effectively and comprehensively.

It's important to note that both projects and their combination were not originally intended for certification by their creators. However, they serve well for this academic exercise, as the focus is on the certification procedure rather than the system's quality or real-world applicability.

4.2 Brief Overview of the EmoPy Framework

The EmoPy framework provides multiple neural network architectures for facial expression recognition, including ConvolutionalNN, TransferLearningNN, and ConvolutionalLstmNN. These architectures vary in complexity, with the ConvolutionalNN being the simplest and TransferLearningNN (using Google's Inception-v3) being the most complex. The authors experimented with different architectures and found that the ConvolutionalNN has the best overall performance (Perez, 2018a). The EmoPy README file suggests using two publicly available datasets for training and evaluation: the Microsoft FER2013 dataset and the Extended Cohn-Kanade dataset. The FER2013 dataset contains over 35,000 facial expression images across 7 emotion classes, while the Cohn-Kanade dataset contains 327 facial expression sequences. To increase the size and suitability of the training and validation datasets, the authors used data augmentation techniques (Angelica Perez et al., 2021).

The neural networks in EmoPy are trained on the training set and evaluated on a separate validation set. The training process involves iteratively adjusting the network weights to minimize the loss between the predicted and labelled emotions. First, the training process involves splitting the dataset into training and validation sets. To mitigate overfitting, the authors monitored the gap between training and validation accuracy. This approach aims to avoid overfitting and ensure the model generalizes well by using validation data not seen during training. Performance is measured using training and validation accuracy. Confusion matrices are also used, which visually represent misclassification rates to help refine the models. Cross-validation against multiple datasets ensures the model's generalizability.

The authors tested the neural network architectures on various emotion classification tasks and reported their performance. The ConvolutionalNN model was found to be the best performer, achieving over 90% accuracy on some emotion subsets. Confusion matrices were also used to gain insight as to how to improve the models (Perez, 2018a). The key goals of the EmoPy project are to provide free, open-source, and easy-to-use facial expression recognition capabilities, and to advance research in this field by making the models and datasets publicly available. All the sources used are collected in the Appendix: The Facial Emotion Recognition System.

4.3 Brief Overview of the RIOT Project

RIOT is a live-action film that dynamically responds to emotions, utilizing facial emotion recognition technology to guide viewers through an ongoing dangerous



Figure 4.1: **RIOT Installation in New York 2018.** This image shows the RIOT Art installation in New York City in 2018. One participant is standing in front of the screen watching the experience. (Perez, 2018a)

riot. The experience allows the audience's emotions to drive the narrative of the film in real-time (Thoughtworks, 2018a). The project began during artist Karen Palmer's 2017-2018 residency at Thoughtworks, where a new iteration of the emotion analysis engine and RIOT user experience was developed. The RIOT installation has been showcased at various events and festivals.

The RIOT experience uses the EmoPy framework and emotion recognition model, that was trained with it, in its emotion recognition system to respond to the participant's emotional state during the live-action film sequence. The characters and narrative adapt based on the viewer's detected emotions, creating an immersive, multisensory experience that enhances the player's cognitive skills and self-awareness according to its creators (Palmer, 2016). The System setup can be seen in the Image 4.1. A participant will stand in front of the screen and watch the experience. In different intervals, the mounted webcam will be used to capture the person's face and predict the person's emotion. According to the emotion detected, the film will proceed differently, making the experience interactive.

4.4 The complete AI application

For the purpose of this sample certification, the RIOT installation is used as a complete system context for the certification process. However, the EmoPy framework, which uses machine learning to detect emotions, plays a significant role as the core system in this installation. The focus of this certification process is on certifying the AI system used within the RIOT context. While the surrounding code, setup, and information that make up the entire art installation are relevant because the AI system cannot be certified independently of these components. There are some shortcomings with this approach, as not all the needed information is available. There are some details regarding this in the Approach section. However, this approach still allows for an exploration of the certification process while acknowledging the limitations and academic nature of the exercise. The following sections will delve into the specific certification procedures applied to this facial emotion recognition system. A complete overview of all resources that are used for certification of this system is given in the Appendix: The Facial Emotion Recognition System.

5 Approach

The general approach of this thesis is to apply the Fraunhofer AI Certification Catalogue to an existing AI system in order to explore and evaluate the certification process. The chosen AI system is the facial emotion recognition component of the RIOT art installation, which utilizes the EmoPy framework. This system was selected by the author because it is open-source and seems to be well-documented. It is also integrated into a larger application context, and it is potentially covered by the EU AI Act, as discussed in Chapter 4. The Fraunhofer Catalogue was chosen as the primary certification framework for this study due to its comprehensive nature and full public availability. It provides a structured approach to AI certification, covering multiple dimensions of risks. The certification process, as outlined in the Fraunhofer Catalogue, involves several key steps:

1. **First Step:** Get an overview of the System (AI Profile (PF)) and define the AI-System and the boundaries to the surrounding system
2. **Second Step:** Define the life cycle of the AI application
3. **Main Step:** Get an overview over all the risk dimensions
 - a) **Protection requirements analysis:** Determine which risk dimensions apply
 - b) **Risk Analysis:** For each applicable dimension:
 - i. Risk analysis and objectives
 - ii. Criteria for achieving objectives
 - iii. Measures
 - iv. Overall assessment of a risk area
 - v. Summary of each dimension
 - vi. Cross-dimension assessment
4. Drawing conclusions and making a certification decision based on the success of the cross-dimensional assessment

After conducting the certification process with the Fraunhofer catalogue, conclusions will be drawn regarding the challenges encountered throughout the process. Several key aspects of the certification process will be addressed in this analysis. It will explore the challenges faced in selecting an appropriate AI application for certification. The discussion will then look at how well the Fraunhofer certification process worked for this specific case, highlighting strengths and areas for improvement. Furthermore, a comparative analysis will be presented, examining how the other two introduced catalogues differ from the Fraunhofer approach and how they could potentially enhance or complement the certification process. Lastly, this study will examine the limitations of this approach. Including what the constraints of the chosen AI system and the certification catalogue are, and if the findings are applicable to other AI applications and certification scenarios. This comprehensive analysis will provide valuable insights into the practical implementation of AI certification processes and contribute to the ongoing discourse on AI certification.

5.1 Before the certification process

Before beginning the actual certification process, several preparatory steps were taken. First, the AI application was selected. The GitHub repository of the EmoPy project was forked and work was done to find the correct dependencies to allow for a detailed examination of the codebase. An extensive research phase followed, focusing on both the EmoPy framework and the RIOT installation. This research culminated in the creation of a comprehensive document basis, which is included in the Appendix (The Facial Emotion Recognition System) of this thesis. This basis, that combined elements of both the EmoPy framework and the RIOT installation, was created to facilitate the certification process. This step was necessary to provide a complete system context for certification, as certifying standalone machine learning models can be challenging with the Fraunhofer certification catalogue and would be markedly different to a real-world certification. During this preparation phase, it became apparent that certain information gaps existed in the original documentation. To make the certification process feasible, some assumptions and additional details had to be filled in. These additions were based on reasonable interpretations of the available information and common practices in AI development. These additions were minor in scale and kept to the bare minimum, to be able to do a certification.

5.2 AI Profile

The first formal step in the certification process was to complete the AI Profile, as specified in the Fraunhofer Catalogue. This step was straightforward because of the thorough research conducted in the preparation phase. The AI Profile provided a structured overview of the system's functionality, intended application context, and key characteristics. The completed AI Profile is in the Appendix: AI Profile.

5.3 Life Cycle of the AI Application

Following the AI Profile, a life cycle overview was conducted. Although not explicitly stated as a distinct step in the Fraunhofer Catalogue, the author considered it beneficial to gain a thorough understanding of the AI system's development and operation stages. The AI life cycle is described as all the stages an AI system undergoes, from planning and development to deployment, operation, ongoing maintenance and potentially continued model training, ensuring trustworthiness and compliance throughout its use. The questions for this life cycle overview were adapted from a table in the Fraunhofer Catalogue, covering aspects such as data acquisition, model development, and operational considerations. The AI life cycle analysis is in the Appendix: AI life cycle.

5.4 Protection Requirement Analysis

The protection requirement analysis is an important first step in the certification process, as it identifies the risk dimensions that require more in-depth analysis. This analysis involves evaluating the potential impact of the AI system across various dimensions such as fairness, reliability, and data protection. The analysis can be seen in the Appendix: Protection Requirement Analysis. All dimensions were examined and multiple were found, that are of medium or high risk. For the purposes of this work, only two of the required risk dimensions were selected for detailed risk analysis: reliability and fairness. This selection was made to focus the study and manage the scope of the certification process within the constraints of a bachelor thesis. Exploring those two dimensions is also sufficient to understand the certification procedure and draw according conclusions.

5.5 Risk Analysis

The risk analysis forms the core of the certification process. This step consisted of working through a questionnaire provided in the Fraunhofer Catalogue. The questions addressed different aspects of the selected risk dimensions. For each dimension, the analysis covered topics such as data quality, model design, testing procedures, and operational considerations. Following the individual dimension analyses, a cross-dimensional assessment was conducted to identify any potential trade-offs or interactions between the examined dimensions. This step is important to ensure a complete understanding of the AI system's performance and risks. This can be seen in the Appendix: Risk Analysis.

The challenges encountered during this certification process are described in detail in the next chapter, Findings. Based on these experiences, conclusions are drawn. Additionally, comparisons with two other certification catalogues are discussed to provide a broader perspective on AI certification methodologies.

6 Findings

The certification of the chosen facial emotion recognition system was performed, and the results can be seen in the Appendix. The core of the certification with the Fraunhofer Catalogue is the Protection Requirement Analysis and the Cross-dimensional assessment. Performing the certification allows this thesis to highlight challenges that arise from using this certification approach in general. But it also shines a light on issues that make a work such as this one more difficult, such as finding a suitable AI system to certify in the first place. This work did not prove that the system is compliant with today's regulations, it also never intended to. The certification itself did not look at all necessary dimensions to be able to fully certify the system. The conclusion of the certification suggests, that not even all the dimensions that were closely examined are sufficient for a certification (Risk Analysis). In a real certification scenario, these shortcomings would be communicated to the development team, to address the issues and allow the system to be certified. If full certification were feasible, the process would demonstrate the AI system's compliance with specific standards that are outlined in the Fraunhofer Catalogue. All the challenges and potential improvements that were found are described in the following paragraphs.

6.1 Selecting the AI-System

In a conventional certification scenario, the process of selecting an AI system for certification is usually not a consideration, as the system to be certified is predetermined by the organization seeking certification. However, for the purposes of this work, the selection of an appropriate AI system represented a crucial first step that significantly influenced the subsequent certification process and what can potentially be learned from it. The selected system provided a mostly robust foundation for the certification effort. Its existing application context, and good documentation of both the AI model and its surrounding system, enabled meaningful progress through the certification process. During the certification process, the importance of considering both technical factors and solid documentation when choosing an AI system for certification became evident.

6.1.1 Initial Considerations and Challenges

The selection of an appropriate AI system requires careful consideration of multiple interconnected factors. While an initial approach might suggest identifying and selecting a standalone AI model or neural network, this proves insufficient when considering the comprehensive requirements of a certification processes. Particularly, the Fraunhofer catalogue that was primarily used in this thesis, takes an extensive look at how to define the system exactly and how to define the boundary to other software components. This cannot be done with a standalone AI model.

6.1.2 Context and Embedding Requirements

The certification process inherently demands a broader contextual framework than what might be immediately apparent. Rather than existing in isolation, the AI system must be embedded within a larger operational context and demonstrate clear use case applications. While it would be theoretically possible to construct artificial use cases for the certification purpose, such an approach could result in a suboptimal certification scenario. In this research, the selection process benefited from identifying a system that already possessed an established real-world application context. This characteristic proved invaluable, as it enabled a more natural translation into a certifiable system, providing the necessary surrounding information to support a comprehensive certification approach. The existence of this practical context significantly enhanced the certification process's authenticity and relevance.

6.1.3 Documentation Requirements

Documentation emerges as another critical factor in the selection process, operating on two distinct levels. First, the AI model itself must be thoroughly documented, providing technical specifications and operational parameters. Second, and equally important, the surrounding system infrastructure must be comprehensively documented to make certification feasible. This documentation requirement significantly narrows the field of suitable candidates for certification studies. A particular challenge encountered in this work relates to the absence of a development team. When selecting an existing model for certification, there is typically no active development team invested in the certification process. This situation creates a significant constraint. It is not possible to request additional documentation or engage in an iterative process with developers to enhance the existing documentation. Consequently, the available documentation must be sufficiently

comprehensive from the outset, as it is not possible, to supplement or clarify information gaps that might emerge during the certification process.

6.2 The Certification Process

The Fraunhofer catalogue was used as the primary basis for this certification due to its comprehensive and detailed nature. While other catalogues were considered, such as the TÜV catalogue, they presented significant limitations. The TÜV catalogue's incomplete publication made it unsuitable to use for certification. The Auditing Machine Learning Algorithms catalogue, while fully published and potentially suitable for certification, employs a substantially different approach compared to the Fraunhofer catalogue. Its less step-by-step nature potentially presents additional challenges for those with limited certification experience.

6.2.1 Challenges During Certification

The system's documentation

The certification process encountered several key challenges. One fundamental challenge is, that the system's documentation was never originally intended for certification purposes. The nature of the development also never called for extensive and detailed documentation. This limitation created occasional gaps in documentation that would have been essential for a complete certification. In some instances, adequate substitutions were made beforehand to have a more realistic certification scenario. The documentation of a system is key for a certification. Choosing a publicly available system, that was not intended for certification, has its shortcoming. This makes a sample certification, such as the one that has been attempted here, more difficult and potentially less meaningful.

No active development team

The absence of an active development team emerged as a critical limitation in the certification process. Without ongoing development support, it was impossible to implement the typical feedback loop where certification findings would normally lead to documentation improvements and system adjustments. In a standard certification scenario, identified gaps or shortcomings can trigger an iterative process of enhancement, with the development team actively working to make the system more certifiable. However, in this case, the system had to be evaluated purely on its existing documentation and state. Therefore, there would have been only

two possible outcomes: either certifiable or not certifiable with the available materials. This limitation was further complicated by the fact, that the system was originally developed several years ago, and the entire development team moved on from the project. The lead developer generously provided time to answer questions. But because the project is old, this means that certain details had become less accessible or clear over time. This combination of inactive development and the system being old created a static evaluation scenario, rather than the dynamic, iterative process that typically characterizes successful certification efforts where development teams actively work towards certification compliance.

6.2.2 Specific Observations on the Fraunhofer Catalog

The implementation of the Fraunhofer catalogue revealed several notable characteristics and challenges. The catalogue’s documentation-centric approach makes it nearly impossible to use for code-only projects, as it focuses exclusively on documentation rather than direct code examination. While code can inform the certification process and documentation creation, the catalogue never directly addresses or describes code. The catalogue’s high specificity and detail provide comprehensive coverage, reducing the likelihood of overlooking critical aspects. However, this thoroughness occasionally results in similar or nearly duplicate questions, increasing the time required for certification completion. The strong documentation focus means less direct attention to mathematical or technical system operations. While the neural network structure remains important for documentation purposes, it is examined only implicitly rather than explicitly. This approach can be advantageous when dealing with proprietary information, as documentation alone might suffice for a potential certification. A particular strength of the Fraunhofer catalogue lies in its clear differentiation between the AI model, system, and embedding code, which proves crucial in determining certification scope and requirements. This distinction helps ensure appropriate certification coverage.

TÜV Catalogue

One notable limitation of the Fraunhofer catalogue is the lack of guidance on how to answer the posed questions. In this regard, a more technology-centric catalogue like TÜV’s could provide valuable complementary guidance. The TÜV catalogue, despite its publication limitations and restricted focus on machine learning and supervised learning systems, offers useful insights into the technical aspects of ML systems operations.

Auditing Machine Learning Algorithms Catalogue

The Auditing Machine Learning Algorithms Catalogue presents a markedly different structural approach compared to Fraunhofer's. Its topic-based organization consolidates related questions – for instance, grouping all data-related questions together – contrasting with Fraunhofer's distributed approach where data-related questions appear across various subsections. This structural difference complicates the potential combination of these catalogues. However, the auditing catalogue's reduced duplication could potentially streamline the certification process.

6.3 Learnings

The certification process revealed several significant insights regarding both methodological approaches and practical certification challenges.

6.3.1 Catalogue-Specific Observations

The Fraunhofer catalogue, while demonstrating robust effectiveness, revealed both strengths and limitations in practical application. Its exhaustive and detailed nature ensures comprehensive coverage but is time intensive. This thoroughness, while beneficial for certification rigour, needs to be balanced against practical time constraints in real-world scenarios. The evaluation of alternative catalogues provided additional insights. The TÜV catalogue's incomplete publication status rendered it unsuitable for standalone certification efforts. The Auditing Machine Learning Algorithms catalogue showed promise for certification purposes, potentially offering a more streamlined approach compared to the Fraunhofer methodology. However, its less structured nature suggests a need for deeper AI system expertise. But it might potentially be a faster certification process while maintaining quality standards.

6.3.2 Limitations

Several key limitations emerged during the certification process. The absence of an active development team is limiting, as it prevented the implementation of the typical feedback loop essential for certification refinement. This limitation transformed the certification process into evaluation rather than an iterative improvement process, highlighting the importance of ongoing development support for successful certification efforts. Documentation gaps could not be addressed through subsequent submissions. This emphasized the importance of comprehensive initial documentation of the chosen system.

Any potential future sample certification of systems that are not actively developed should keep this in mind and a system with the most comprehensive documentation should be chosen.

6.3.3 Recommendations for future Sample Certifications

The experience gained from this study suggests several crucial considerations for future certification efforts. For model selection, emphasis should be placed on identifying AI systems that exist within a broader application setting with surrounding code. The AI system itself, as well as the application setting and surrounding code, should be extensively documented. The ideal certification candidate should have an active development team willing to engage in the certification process.

A practical recommendation that emerged from this study was the value of creating a centralized archive of all available information and documentation before initiating the certification process.

6.3.4 Recommendations for Real-World Applications

The findings from this research yield several practical recommendations for real-world certification implementations. The Fraunhofer catalogue, while highly detailed and extensive, requires significant time investment for thorough completion. However, its precision and comprehensiveness make it a valuable tool for certification processes. Particularly noteworthy are the initial sections of the catalogue, specifically the AI lifecycle overview, which prove especially effective in providing certifiers with comprehensive insights into the AI system's general functionality. This initial overview serves as an excellent starting point for any certification process. While the Auditing Machine Learning Algorithms Catalogue was not extensively examined in this work, its different structural approach suggests potential for more efficient certification processes, potentially offering certifiers a faster way to system certification while maintaining appropriate quality.

7 Conclusion

This research into the practical application of AI certification catalogues yielded significant insights regarding both the capabilities and limitations of current certification approaches, and what the shortcomings of certifying an open-source AI system are. The implementation of the Fraunhofer AI Assessment Catalogue demonstrated its effectiveness as a comprehensive certification tool, particularly in its systematic approach to evaluate AI systems. It also suggests, that the approach is at times bulky and time-consuming in some areas. Other approaches and certification catalogues could be considered, to potentially streamline the process. The certification process highlights the critical importance of complete system documentation and active development team engagement. The absence of these elements can negatively impact the certification itself and can lead to certification infeasibility. During the certification attempt, it became clear which characteristics an AI system must possess to be able to adequately certify a system. Before selecting a public AI system for a sample certification, consideration should be given to factors like thorough documentation and accessibility to the development team. These findings directly addressed the initial research objectives by identifying useful aspects of this certification approach and some that could be improved. It also provides practical insights into the certification process and its limitations. Future work in this field could focus on developing more flexible certification methodologies that can accommodate various system states and development scenarios, while allowing for a faster certification process.

Bibliography

- C. J. Costa and M. Aparicio, “Applications of Data Science and Artificial Intelligence,” *Applied Sciences*, vol. 13, no. 15, p. 9015, Aug. 2023, number: 15 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: <https://www.mdpi.com/2076-3417/13/15/9015>
- M. Kasinidou, S. Kleanthous, M. Busso, M. Rodas, J. Otterbacher, and F. Giunchiglia, “Artificial Intelligence in Everyday Life 2.0: Educating University Students from Different Majors,” in *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*, ser. ITiCSE 2024. New York, NY, USA: Association for Computing Machinery, Jul. 2024, pp. 24–30. [Online]. Available: <https://dl.acm.org/doi/10.1145/3649217.3653542>
- K. Baum, J. Bryson, F. Dignum, V. Dignum, M. Grobelnik, H. Hoos, M. Irgens, P. Lukowicz, C. Muller, F. Rossi, J. Shawe-Taylor, A. Theodorou, and R. Vinuesa, “From fear to action: AI governance and opportunities for all,” *Frontiers in Computer Science*, vol. 5, May 2023, publisher: Frontiers. [Online]. Available: <https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2023.1210421/full>
- N. A. Smuha, “From a ‘race to AI’ to a ‘race to AI regulation’: regulatory competition for artificial intelligence,” *Law, Innovation and Technology*, vol. 13, no. 1, pp. 57–84, Mar. 2021, publisher: Routledge _eprint: <https://doi.org/10.1080/17579961.2021.1898300>. [Online]. Available: <https://doi.org/10.1080/17579961.2021.1898300>
- M. Mueck, S. Cadzow, C. Communications, and S. Wood, “ETSI Activities in the field of Artificial Intelligence Preparing the implementation of the European AI Act - 1st Edition – December -2022,” ETSI, Tech. Rep., 2022.
- T. Nad, S. Scher, and F. Königstorfer, “Trustworthiness of AI,” SGS, Tech. Rep. NIST AI 100-1, Jun. 2023.
- P. M. Winter, S. Eder, J. Weissenböck, C. Schwald, T. Doms, T. Vogt, S. Hochreiter, and B. Nessler, “Trusted Artificial Intelligence: Towards Certification of

Bibliography

- Machine Learning Applications,” Mar. 2021, arXiv:2103.16910 [cs, stat]. [Online]. Available: <http://arxiv.org/abs/2103.16910>
- D. M. Poretschkin, A. Schmitz, D. M. Akila, L. Adilova, D. D. Becker, D. A. B. Creemers, D. D. Hecker, D. S. Houben, J. Rosenzweig, J. Sicking, E. Schulz, D. A. Voss, and D. S. Wrobel, “AI Assessment Catalog,” Fraunhofer IAIS, Tech. Rep., Feb. 2023.
- B. Pimentel, “Why AI still needs regulation despite impact,” Feb. 2024. [Online]. Available: <https://legal.thomsonreuters.com/blog/why-ai-still-needs-regulation-despite-impact/>
- S. Benjamens, P. Dhunoo, and B. Mesko, “The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database,” *npj Digital Medicine*, vol. 3, Sep. 2020.
- A. D. Saenz, Z. Harned, O. Banerjee, M. D. Abràmoff, and P. Rajpurkar, “Autonomous AI systems in the face of liability, regulations and costs,” *npj Digital Medicine*, vol. 6, no. 1, pp. 1–3, Oct. 2023, publisher: Nature Publishing Group. [Online]. Available: <https://www.nature.com/articles/s41746-023-00929-1>
- European-Union, “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.” Jun. 2024, legislative Body: CONSIL, EP. [Online]. Available: <http://data.europa.eu/eli/reg/2024/1689/oj/eng>
- F. Fernhout and T. Duquin, “The EU Artificial Intelligence Act: our 16 key takeaways | Stibbe,” Feb. 2024. [Online]. Available: <https://www.stibbe.com/publications-and-insights/the-eu-artificial-intelligence-act-our-16-key-takeaways>
- T. Madiaga, “Artificial intelligence liability directive,” Jan. 2023.
- European-Union, “Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive),” 2022. [Online]. Available: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0496>

Bibliography

- White-House, “Blueprint for an AI Bill of Rights,” Jan. 2022. [Online]. Available: <https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf>
- M. . S. Department for Digital, Culture, “Establishing a pro-innovation approach to regulating AI,” Jul. 2022. [Online]. Available: <https://www.gov.uk/government/publications/establishing-a-pro-innovation-approach-to-regulating-ai/establishing-a-pro-innovation-approach-to-regulating-ai-policy-statement>
- M. Blösser and A. Weihrauch, “A consumer perspective of AI certification – the current certification landscape, consumer approval and directions for future research,” *European Journal of Marketing*, vol. 58, no. 2, pp. 441–470, Jan. 2023, publisher: Emerald Publishing Limited. [Online]. Available: <https://doi.org/10.1108/EJM-01-2023-0009>
- R. Delgado-Aguilera Jurado, X. Ye, V. Ortola Plaza, M. Zamarreño Suárez, F. Pérez Moreno, and R. M. Arnaldo Valdés, “An introduction to the current state of standardization and certification on military AI applications,” *Journal of Air Transport Management*, vol. 121, p. 102685, Sep. 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0969699724001509>
- Benjamin Fresz, Vincent Philipp Göbels, Safa Omri, and Danilo Brajovic, “The Contribution of XAI for the Safe Development and Certification of AI: An Expert-Based Analysis,” Jul. 2024. [Online]. Available: <https://arxiv.org/html/2408.02379v1>
- Hadrien Pouget, “Standardsetzung | EU-Gesetz zur künstlichen Intelligenz,” 2024. [Online]. Available: <https://artificialintelligenceact.eu/de/standardeinstellung/>
- SAI-FI-DE-NL-NO-UK, “Auditing machine learning algorithms,” Supreme Audit Institutions FI, DE, NL, NO, UK, Tech. Rep., Feb. 2023.
- Angelica Perez, Julien Deswaef, and Puneetha Pai, “thoughtworksarts/EmoPy,” Jan. 2021, original-date: 2017-12-20T02:19:22Z. [Online]. Available: <https://github.com/thoughtworksarts/EmoPy>
- Thoughtworks, “RIOT | Thoughtworks Arts,” 2018. [Online]. Available: <https://thoughtworksarts.io/projects/riot/>
- A. Perez, “EmoPy: a machine learning toolkit for emotional expression,” Sep. 2018. [Online]. Available: <https://www.thoughtworks.com/insights/blog/emopy-machine-learning-toolkit-emotional-expression>

Bibliography

- K. Palmer, "RIOT AI," Mar. 2016. [Online]. Available: <http://karenpalmer.uk/portfolio/riot/>
- Thoughtworks, "thoughtworksarts/riot," Sep. 2019, original-date: 2017-11-22T15:59:17Z. [Online]. Available: <https://github.com/thoughtworksarts/riot>
- TED Residency, "Karen Palmer: The film that watches you back," May 2018. [Online]. Available: <https://www.youtube.com/watch?v=Rw8gLEkFdSw>
- Thoughtworks, "Thoughtworks Arts Exhibition at SPRING/BREAK for Armory Week 2018 | Thoughtworks Arts," Apr. 2018. [Online]. Available: <https://thoughtworksarts.io/blog/thoughtworks-arts-exhibition-spring-break-armory-week/>
- Thoughtworks, "Karen Palmer Awarded Thoughtworks AI Residency | Thoughtworks Arts," Jul. 2017. [Online]. Available: <https://thoughtworksarts.io/blog/karen-palmer-ai-residency/>
- Karen Palmer, "RIOT Video," Mar. 2017. [Online]. Available: <https://www.youtube.com/watch?v=-BCny9SuI3A>
- A. Perez, "Recognizing human facial expressions with machine learning," Aug. 2018. [Online]. Available: <https://www.thoughtworks.com/insights/articles/recognizing-human-facial-expressions-machine-learning>
- E. D. Team, "Welcome to EmoPy's documentation! — EmoPy 1.0 documentation," 2017. [Online]. Available: <https://emopy.readthedocs.io/en/latest/>
- Microsoft, "microsoft/FERPlus," 2023, original-date: 2016-09-14T06:35:21Z. [Online]. Available: <https://github.com/microsoft/FERPlus>
- J. Cohn, "Resources – Jeffrey Cohn," 2024. [Online]. Available: <https://www.jeffcohn.net/resources/>

Appendix

The Facial Emotion Recognition System

This serves as an overview of all the available information and documentation of the System that this work certifies. All the Information that is listed and quoted here is treated as documentation for the certification process. To make the certification feasible, some assumptions and additional details are filled in. These additions are based on reasonable interpretations based on the following documentation.

1 RIOT Installation

- GitHub Repository of the RIOT Art Installation (Thoughtworks, 2019)
- Article on the RIOT Art Installation (Thoughtworks, 2018a)
- TED Talk on the RIOT Art Installation (TED Residency, 2018)
- Article that describes different Art Installations (one of them is the RIOT Art Installation) (Thoughtworks, 2018b)
- Article on Karan Palmer (the artist behind the RIOT Art Installation) (Thoughtworks, 2017)
- Short description of the RIOT Art Installation by Karan Palmer (Palmer, 2016)
- Video that showcases and describes the RIOT Art Installation (Karen Palmer, 2017)

2 EmoPy Framework

- GitHub Repository of the EmoPy Framework (Angelica Perez et al., 2021)

- Article describing the EmoPy Framework and technical decisions that were made in more detail (Perez, 2018a)
- Article describing in more detail decisions that were made regarding the architecture of the Emotion Recognition model (Perez, 2018b)
- Python Documentation of the EmoPy Framework (Team, 2017)
- GitHub Repository of the FER+ Dataset (Microsoft, 2023)
- Extended Cohn-Kanade Dataset (Cohn, 2024)

3 The AI Application that is certified

To make the certification process feasible, these are the assumptions that were made:

- The AI Application is not the entire RIOT Art Installation, but only the system subset that receives centred images from the webcam and returns the emotion predictions
- Since the pretrained model and the target emotions used during training are not disclosed, it is assumed that the EmoPy ConvolutionalNN model is used without any alterations. Furthermore, that the target emotions are anger, fear, calm, surprise. It is presumed that the model with these features that is discussed in one of the articles with the according confusion matrix is the model that is used in the installation.
- All Items that are listed above are treated as if they are proper documentation of the system. Even items that are articles or videos are treated as if they are written usable documentation.

AI Profile

All questions and descriptions of the AI Profile are taken as is from the Fraunhofer Certification Catalogue (Poretschkin et al., 2023). All the answers and descriptions are given by the author of this work and are always written inside the black boxes. The answers provided here are intentionally kept short because the approach of this work doesn't require more detail, making the process more streamlined. In a real certification scenario, however, the answers would typically be longer and more detailed.

Functionality and intended application context (FA) [PF-T-FA-01] Describe the task and/or functionality of the AI application. Also explain the following points while doing this:

- Which problem does the AI application solve? (What exactly does it "do"?)
- What input data is provided and of what type is it?
- What are the outputs of the AI application and of what type are they?

The AI system is a crucial component of the RIOT Installation, responsible for performing emotion recognition and providing the primary classified emotion back to the wider system. It addresses the challenge of automatically identifying the user's emotion at specific moments and intervals during the video playback of the installation. The AI system receives a 48x48 pixel image from a webcam as its input. The webcam is positioned in front of the user, who is illuminated by two lights, and the image is pre-centered by a different software component. The output consists of three recognizable emotions (anger, fear, calm) along with their respective prediction percentages. The emotion with the highest percentage is used, in the wider system, to choose the corresponding video clip.

[PF-T-FA-02] Describe in more detail the intended application context and operating environment of the AI application. Also explain the following points while doing this:

AI Profile

- Is the AI application embedded in a larger surrounding system? If this is the case, describe the relationship and interaction between the AI application and the larger surrounding system. In particular, outline the interfaces.
- To what extent are humans involved in operating or supervising the AI application?

The AI system is used in the RIOT Installation. It is a responsive, live-action film that uses facial expressions to navigate through a film sequence. The installation is in some sort of ART Installation indoors. The user stands in front of a screen with a webcam mounted in front of the user. The users' face is also illuminated by lights. Influence by outside lights or other interferences, that could disturb the AI system are managed, and within specification. The AI system is used within this wider system. The primary interactions are receiving the preformatted image from this wider system and returning the emotions with the according recognised percentages. The emotion with the highest emotion will be used to select the next film clip. Humans are not involved in the system during the film experience. Humans might be involved, showing users where to stand and what to expect. During the experience, the system is completely autonomous.

[PF-T-FA-03] What are the requirements for the AI application in terms of regulatory affairs, economic feasibility and avoiding possible material and non-material risks (e.g., functional safety, IT security, personal rights) in the intended application context?

Given that the AI application plays a crucial role in shaping the user experience, it is vital that the system operates flawlessly to ensure the interaction is not compromised. While the primary risk to the user is minimal – resulting mainly in less accurate emotion recognition that might lead to a different narrative path in the interactive film – the overall experience may be diminished. A more significant concern lies in the potential risks associated with capturing the user's face via the webcam. If this data were to be stored or sold, it could infringe on the user's privacy and rights.

[PF-T-FA-04] In which other application contexts is the AI application conceivable? And in which application contexts or operating environments related to [PF-T-FA-02] should the AI application not be used?

AI Profile

The AI application is conceivable in multiple other applications. Some examples could be emotion recognition by a TV of its audience and maybe using this data to drive advertising. Other applications might be in automotive settings, like to recognise a driver's emotion and maybe warn the driver to take a break from driving. Importantly, the AI system has primarily been tested in situations where the user is centred and looking directly at the camera. Moreover, the impact of light exposure, for example from the sun on the outside, could make the system unreliable and the system should not be used in those situations.

[PF-T-FA-05] Is there any other important information about the functionality of the AI application or its operating environment?

All the important information is already mentioned above.

Structure of the AI application (ST) [PF-T-ST-01] Describe the structure of the AI application. To do this, outline:

- A list of the most important components (AI component, other software modules) and the specification of their functionalities,
- The architecture of the AI application and how the individual components interact with each other.

The AI application has at its core a pretrained Machine Learning model. This model is loaded and used to make the emotion prediction. Keras is used to load and execute the pre-trained machine learning model. The image that the model has to classify is captured by a webcam, the embedding captures the image and makes it available to the AI component. The image is then preprocessed and transformed to the size of 48x48 pixels and the correct dimensions. This image in the form of an Array is given to the ML model and classified. The classification result is given in three percentage values, one for every recognizable emotion. This result is the output of the AI system and is then used by the wider system.

[PF-T-ST-02] Describe the AI component in more detail. In doing so, provide the following information:

- On which ML model or learning algorithm is the AI application based?
- Does the AI component learn in operation continuously, at regular intervals or by initiating retraining?

AI Profile

The ML model was created and trained using the EmoPy framework. A Convolutional Neural Network architecture was used with multiple convolution and polling layers. Multiple other architectures were tested, but the ConcolutionalNN had the best performance, for the given task. The AI component does not learn in operation continuously. It is pretrained and used as such. Using a new and updated model, would require complete retraining and testing by the developers.

[PF-T-ST-03] Are there any other important points about the structure of the AI application?

All the important information is already mentioned above.

AI life cycle

This section should give an overview of the life cycle related risks for the AI system. All questions and descriptions are taken as is from the Fraunhofer Certification Catalogue (Poretschkin et al., 2023). All the answers and descriptions are given by the author of this work and are always written inside the black boxes. The answers provided here are intentionally kept short because the approach of this work doesn't require more detail, making the process more streamlined. In a real certification scenario, however, the answers would typically be longer and more detailed.

What are data related risks, particularly considering the following aspects?

- Acquisition, selection
- Labelling
- Pre-processing (cleansing, enriching, etc.)

The data that is needed for training and validation is critical to the ML model's performance and its characteristics. During development, two datasets were used: the Microsoft FER+ dataset and the Extended Cohn-Kanade dataset. Finding enough open-source data proved challenging, as most large emotion datasets are not publicly available. The provided datasets were already labelled. It is challenging to find data that is both diverse and reflects the user base. The datasets lacked labels for gender, age or ethnicity, which is advantageous as the model cannot use such characteristics to affect predictions unfairly. However, this also complicates testing the system across these particular features. Pre-processing was also used, as not all images in the dataset were used for training. Furthermore, data augmentation was used to great effect, improving the prediction accuracies.

What are related risks during development, particularly considering the following aspects?

- Design of AI component

AI life cycle

- Optimization (training)
- Testing
- Design of embedding

The design of the AI component is essential to achieve high accuracy and reliability. During development, multiple different ML model approaches and structures were used and tested. The Convolutional Neural Network had the best performance was therefore used. The training of the model was done multiple times with different subsets of the datasets and with different target labels. Models that had seven target emotions as output would perform worse, than models that only needed to distinguish between three emotions. As only three target emotions were needed for the wider system, this made finding a suitable model easier. Testing was done with a separate dataset, that was not used during training. An Accuracy measure was used for the training and test data, to see how well the model performs. The accuracy shows how well the neural network predicts emotions from its training data. A confusion matrix was also used to further study misclassifications of the trained models during the model testing. The behaviour of the model in the wider system was also important. Centering the webcam image on the users' face and good lighting are important for the performance.

What are related risks during operation, particularly considering the following aspects?

- Monitoring
- Adjustment/Feedback
- Further learning
- Concept/Model drift

During operation, the system works autonomously. There is no human feedback or check in place, as a human assessment would also be too slow in the film setting to make a prediction or correct the model's decision. If the AI system's prediction is wrong, the interactive experience continuous but on a different story path. This could lead to a degraded experience for the user. The model does not do further learning and is "as is" after the initial implementation. Concept or model drift should not be a factor for this classification task, as human emotions and their corresponding facial features should not change over time. The model also does not change during operation, mitigating these risks.

Protection Requirement Analysis

The tables describing the different risk levels for each dimension are taken as is from the Fraunhofer Certification Catalogue (Poretschkin et al., 2023). All the answers and descriptions are given by the author of this work and are always written inside the black boxes.

1 Dimension: Fairness

High	The AI application controls access to essential services/activities or makes decisions that have a wide-ranging impact on personal rights. Examples: Granting of a visa, admission to schools/universities, automated credit lending, decision on the type of medical treatment
Medium	The output of the AI application is related to a person, even if only in a broader sense. This includes both AI applications that output a decision about a person or categorize a person and AI applications that process person-specific inputs (e.g., speech recognition that transcribes the user's spoken sentences into text). The output of the AI application is not sensitive and does not have any significant impact regarding the personal rights of the people affected. Examples: Recommendation for facial recognition on photos on social media, classification of a person's age based on a photo, speech recognition systems
Low	The AI application does not process any personal data that provides information about age, gender, sexual identity, religion, ideology, ethnic origin or about a possible disability. Furthermore, the function/output of the AI application is not integrated into a process or decision that affects the courses of action or the personal rights of the individual people affected. Examples: Recommendations of personalized advertising, predictions of machine failures

The AI application does process personal information in the form of images of the user's face and recognizes the user's emotion. The output therefore is related to a person in a broader sense. However, the output of the AI application does not have a significant impact on the user's rights. This leads to the assessment of: Medium Risk.

2 Dimension: Autonomy and Control

High	<p>The AI application has a high protection requirement with respect to this dimension if it</p> <ul style="list-style-type: none">• strongly influences the perception or actions of users or subjects over long periods of time or at a risk level that is unacceptable.• limits the ability of users or affected persons to perceive a situation or take action. <p>Examples: An AI application that modulates the voice of care staff so that patients with dementia think they are talking to a relative. An autonomous vehicle that also transports people. An AI application involved in managing access to education by, for example, making admission decisions for attending a university and does not inform the people affected about the use of the AI application.</p>
Medium	<p>The AI application can strongly influence the perception or actions of users or affected persons only temporarily and only under acceptable risk. Examples: An AI application that records the user's fitness, health and nutrition habits and provides them with suggestions and guidelines. An AI application in the form of a doll that simulates human interaction through speech input and output and also facial expressions and movement. An AI application that automatically determines the user's preferences based on previous reading habits and generates a personalized news stream from online content.</p>

Low	The AI application has little influence on the perception or actions of users or affected persons. Examples: An AI application that recognizes bird calls or identifies plants. An AI application for customized route planning. An AI application for customized planning of tourist activities.
-----	--

The AI application does have an influence on the users' perception. The influence is limited to the story path of the film, but essential to the specific experience the user has in the Installation. Therefore, the AI system can strongly influence the perception temporarily. This leads to the assessment: Medium Risk.

3 Dimension: Transparency

High	<p>Non-fulfillment of the transparency requirement (explainability, interpretability or traceability/ reproducibility)</p> <ul style="list-style-type: none">• would either render the AI application useless for its originally intended purpose, e.g., because safe or responsible use does not seem possible,• or would mean that the AI application could only be operated appropriately if additional expenditure is made (unjustifiable in terms of time or money). <p>or would mean that the AI application could only be operated appropriately if additional expenditure is made (unjustifiable in terms of time or money).</p> <p>In addition, a high damage potential already exists if the lack of transparency would lead to a breach of relevant (legal/normative) guidelines. Example: An AI application that makes a medical diagnosis but the decision is not traceable.</p>
------	---

Medium	Situations may arise where failure to meet a transparency requirement lowers the usefulness of the AI application and expenditure (time or financial) is required to ensure the AI application is useful/beneficial. At the same time, a lack of transparency cannot violate relevant (legal/normative) requirements. Example: An AI application used by a company for customer queries (e.g., credit facility queries). If there is an inquiry about the reasons for a specific output, a lack of transparency in the model makes answering it in a satisfactory manner more difficult from the customer's perspective. Example: An AI application that automatically evaluates images on social networks for inappropriate content and blocks the post if necessary. If images are blocked without marking the areas in the image that are crucial to the output or categorizing the alleged violation, this increases the time required for human operators to identify and recheck false positives or false negatives.
Low	There is no transparency aspect that, if it was not fulfilled, could reduce the safety or usefulness of the AI application; or only minor effects on the usefulness of the AI application are possible, which can in turn be fixed with little effort. At the same time, a lack of transparency cannot violate relevant (legal/normative) requirements. Example: An AI-based access control system that uses camera-based facial recognition to determine people's access rights. If these rights are incorrectly classified, specifying the facial features that resulted in the wrong decision would not be useful, as it would not provide an expert with any added value in correcting the decision. In fact, a human supervisor will use their visual system and implicit or explicit professional expertise of personal identification to check the AI application's decision.

The safety and usefulness of the application is not degraded by its transparency. For example: Showing specifically why an emotion is detected, does not enhance the experience or would help it be better when misclassified. This is similar to the example given, where specifying facial features in a face recognition system would not be useful. Therefore, the assessment is: Low Risk.

4 Dimension: Reliability

High	An incorrect prediction of the AI application can lead to bodily injury or major financial damage. Examples: Pedestrian detection in autonomous vehicles, automated credit lending decisions, medical treatment recommendations
Medium	An incorrect prediction of the AI application can at worst lead to medium financial damage. Examples: A poor-quality route planner causes increased energy consumption and a longer time of travel, a faulty forecast of machine utilization in a manufacturing facility causes delays, a defective obstacle detection of a robot vacuum cleaner causes property damage

An incorrect prediction of the AI system does not cause major financial damage or bodily injury. The incorrect prediction does, however, cause a degraded user experience. This might lead to the user not getting an adequate experience. The assessment for the reliability dimension is therefore: Medium Risk.

5 Dimension: Safety and Security

High	<p>At least one of the following applies:</p> <ul style="list-style-type: none">• The AI application interacts with people in such a way that they can be injured if it malfunctions.• A malfunction of the AI application (caused by errors, failures, manipulation or attacks) can result in very high financial damage (e.g., due to property damage). <p>Example: An AI application used to detect people and objects in an autonomous vehicle. Incorrect functioning can result in injury to people and high financial costs due to damage to property. Example: An AI diagnostic application that makes decisions about the type of medical treatment given to individuals. Manipulation of the AI application can result in incorrect treatment of patients and have a serious impact on their health as a result.</p>
------	--

Protection Requirement Analysis

Medium	The AI application cannot cause direct physical injury to people. However, a malfunction of the AI application (caused by errors, failures, manipulation or attacks) can result in high financial damage. Example: An AI-supported application for transporting goods in storage can damage goods, for example, if goods are unintentionally unloaded in areas or at heights where this is not allowed.
Low	A malfunction of the AI application can at worst lead to medium financial damage. Example: An AI application that is used to compose pieces of music. In case of failures or incorrect functioning, no financial damage is expected.

The misclassification of an emotion is the worst malfunction this AI application can have. The maximal financial damage is low, in the form of maybe the users' entry fee, that the user could have paid to be able to experience the installation. Harm to the user beyond this is very unlikely. Therefore, the risk Level is: Low Risk.

6 Dimension: Data Protection

High	The protection requirement is classified as high if one of the following three scenarios applies: Personal data is processed that contains particularly sensitive personal information, or disclosing it would have economic or security-critical consequences for the person in question. Example: Patient file, certificate of good conduct, account information, application documents Licensed data is processed for which the disclosure/access by third parties would violate contractual agreements. Example: Data from other companies was purchased to train the model. Access to this data by third parties would violate the contractual agreements. Organization/business-related data is processed which, if it became known/was accessed by third parties, would severely damage the integrity or competitiveness of the organization. Example: Model extraction would mean that the corresponding AI application could be copied or deliberately manipulated by other organizations.
------	--

Protection Requirement Analysis

Medium	The protection requirement is evaluated as medium if one of the following three scenarios applies and there is no high potential for damage for any of the named data categories (personal/business-related or licensed) according to the top row of this table. The AI application only processes/stores data that does not contain sensitive personal information or that would not cause a major economic disadvantage or threaten the security of the subject if accessed by a third party. The AI application processes/stores licensed data, which when accessed by third parties could result in negligible consequences. The AI application processes/stores business-related data, the disclosure of which could result in medium economic damage that does not threaten the existence of the company. Example: Leisure interests of a person, played tracks, videos viewed in anonymized form Example: An AI application that performs trend analysis based on publicly available social media data.
Low	The AI application does not request, process or store personal data. In addition, the AI application does not store/process any licensed data. Disclosure of the processed data and model characteristics (e.g., model parameters) would have no or negligible impact on the integrity or competitiveness of the organization. Example: A company uses a standard AI solution to predict market development. Other companies in the industry have similar solutions and it is assumed that there is no incentive on the part of the competition to expose or copy this system. For example, data from the DAX or other economic indicators that are freely available are used.

The AI application does need personal data in the form of a live image feed of the user's face. It does fall under one of the categories mentioned at the medium level. The application does not use or store sensitive personal data; however, it uses personal data to function. There is no high potential for damage in this data domain. Therefore, the assessment is: Medium Risk.

Risk Analysis

All questions and descriptions are taken as is from the Fraunhofer Certification Catalogue (Poretschkin et al., 2023). All the answers and descriptions are given by the author of this work and are always written inside the black boxes. The answers provided here are intentionally kept short because the approach of this work doesn't require more detail, making the process more streamlined. In a real certification scenario, however, the answers would typically be longer and more detailed.

1 Dimension: Reliability

1.1 Reliability in standard cases

Risk analysis and objectives

Determining the application area and risk assessment

Application Domain

The application domain is defined. The input data that is required is specified. This, however, is not described in great detail. There are also no examples given that describe the application domain. But the application is of small scale. Therefore, the given information is sufficient.

Risk analysis

There was an assessment made as to what happens when other input than the inspected is received. The emphasis was on the inadequate illumination for the user. Other aspects are not discussed in detail, such as a failure of the webcam. There was no assessment of the resulting risk. However, the risk can be assumed to be low in this application. All in all, that makes this point sufficiently fulfilled.

Objectives

There is no documentation for residual risks, and there are no objectives set. That makes this point not fulfilled.

Criteria for achieving objectives

Quantification of reliability

The documentation covers which performance metrics should be used. It is also described, as to why the metric was chosen. The choice of the loss function is not described. Target intervals are also not, defined. That makes this point only insufficiently fulfilled.

Quantification of the application domain coverage

The coverage area of the training data is discussed. There are no metrics used, but the topic is sufficiently discussed.

Quality of training and test data

The data that is used for training is discussed. Its shortcomings are mentioned. A publicly available and well regarded data source is used. There is some investigation into how accurate the labels are. However, as the dataset was compiled by a third party, this cannot be thoroughly verified. The data that is used can be assumed to be of good enough quality. It also fits the requirements that the wider system has. This leads to the risk level being acceptable.

Measures

Data

Origin and quality of the database

The training and test data that is used, stems from public and well regarded datasets. The data is pre annotated and the process of annotation is described in the dataset's descriptions. The suitability of the data and its potential shortcomings are discussed. A discussion of the data's compatibility is also discussed. The data's structure is also discussed and potential shortcomings mitigated. There are measures taken to prevent data leakage. This makes this topic sufficiently discussed.

Choice of database

The choice of the training and test data is discussed. The coverage of the application domain is documented. There were also measures taken to improve the coverage. The training and test data does not stem from the same setup, however, is comparable and suitable for the application. This makes this topic sufficiently discussed.

AI component

Component design choice

There is documentation provided that describes the choice of the model. There is also justification as to why the specific architecture was selected. Other architectures were also examined, the reasons for choosing the final model are described. An explanation of what features were selected is also given. The generalizability of the model is sufficiently investigated. This makes this topic sufficiently discussed.

Systematic search for weaknesses

There was an investigation for weaknesses of the model. The identified weaknesses are described and measures preventing them implemented. This makes this topic sufficiently discussed.

AI component reliability tests

Test of the AI component with new data were performed. These tests are sufficient to prove good robustness.

Embedding

AI application real-world tests

There were tests of the AI component and the wider system; however, the documentation is lacking. That makes this point not fulfilled.

Measures for operation

Supplement to open-world coverage

The AI application is not used in an open world environment. The environment is very well specified and does not need further discussion here. This makes this topic sufficiently discussed.

Overall assessment

Overall assessment

There were no initial target intervals set. However, the AI application is small in scale and all possibilities have been accounted for. The quality of the prediction is, however, sufficiently discussed. Furthermore, the data has the required quality. The overall system metrics suggest that the system works in practice. Overall, the system works sufficiently well for its application.

1.2 Robustness

Risk analysis and objectives

Risk assessment and definition of the application boundary

Risk assessment

Disturbances that can occur are examined. The environmental conditions are narrow due to their setup in the wider system. The risks of misclassification are in line. This makes this topic sufficiently discussed.

Defining the application boundary

The application boundary is clearly defined. The AI system knows what input to expect, and all input it can receive is within specification. This makes this topic sufficiently discussed.

Objectives

There are no objectives set and there is no documentation about this. However, the danger of misclassification is low. Still, this makes this topic insufficiently discussed.

Criteria for achieving objectives

Quantification of the application boundary

The application boundary is described sufficiently. Disturbance levels are not a big issue in this setup. The specification of the input data is specified. This also clearly defines the application boundary. This makes this topic sufficiently discussed.

Quantification of robustness

The robustness of the AI application within its boundary is sufficiently explained. Statistical metrics are used. The documentation used to access robustness is good. There are no intervals specified as targets. The system as whole still performs within expectations. This makes this topic sufficiently discussed.

Coverage of the application boundary

The coverage of the application boundary is described. The boundary is set narrowly, making the application perform well within these limits. This also makes the application less risky within these boundaries. There are no target intervals specified. However, this is not deemed necessary, as the application is narrow in scope. This makes this topic sufficiently discussed.

Measures

Data

Data for testing robustness

The choice of the dataset is discussed. As the application boarder is narrow, the dataset is well suited to cover the required area. Data augmentation was used during development, the issue of overfitting was also looked into. The results were deemed sufficient for the application. The documentation of the data assessment was already discussed and deemed sufficient. This makes this topic sufficiently discussed.

Data for robust training

There was no additional (special) data used to increase robustness. However, the entirety of the dataset was chosen to achieve a high enough classification accuracy. This makes the dataset that was used sufficient. Augmentation was also used to achieve this. The data was examined and issues with representativeness were found. However, methods were found to increase representativeness up to an acceptable level. This makes this topic sufficiently discussed.

Examining corner cases

There is no documentation that examines corner cases. Some discussion in this area happened. However, this is an area that was not sufficiently looked into. This makes this topic insufficiently discussed.

AI component

Development and training procedure

As the possibility for disturbances is low, and the remaining disturbance factors were discussed, the model training was adapted to make the ML model robust. The used architecture proofed sufficient for the task. It generalized well enough to be used in the application. This makes this topic sufficiently discussed.

Testing of AI component robustness

Test of the AI model were performed to some extent. However, the documentation is lacking. There also were no previously set target intervals. This makes this topic insufficiently discussed.

Embedding

Real-world generalization/exploration testing of the AI application

There is no documentation on real-world generalization testing. This makes this topic only insufficiently discussed.

Measures for operation

Monitoring input data in operation

As the input data is in a standard format, and well described, there are no special systems in place to monitor the input data in operation. This, however, is not deemed necessary. This makes this topic insufficiently discussed.

Monitoring outputs in operation

The application is happening in real time, making monitoring the output operation not possible by humans. There is no redundant system that checks the output. But this is also not deemed necessary. Due to the low risk of the application, this topic is not applicable.

Overall assessment

Overall assessment

The requirements in this area are met in general. There are some shortcomings in the documentation. As the application is low risk and the application is narrowly defined, and the input is also narrowly defined. The provided documentation is still sufficient.

1.3 Interception errors at model level

Risk analysis and objectives

Scope, risk analysis and objectives

Scope

Input that is outside the application domain is not likely in this setup. There would be only minor additional risk. Identification of possible out of spec input is not required.

Risk analysis

The risk of a situation with out-of-spec input is minor. Therefore, there was not much attention given to this sort of error.

Objectives

As there were no particular dangers identified, there are also no objectives set.

Criteria for achieving objectives

Out-of-distribution coverage

/

Existence of mitigation strategies

/

Requirements for the detection methods

/

Measures

Data

Out-of-distribution data set

/

Data set splits for extrapolation

/

AI component

Design for intercepting errors in outputs with correlation-based methods

/

OOD tests

/

Extrapolation test

/

Uncertainty estimation

/

Embedding

Real-world tests

/

Measures for operation

Monitoring of input and output data

/

Overall assessment

Overall assessment

As the possible errors and possible negative effects are deemed very low for this area, no in detail analysis was done. This area and possible dangers are covered by other areas such as robustness. The area intercepting errors at model level is therefore deemed not necessary for this application.

1.4 Uncertainty estimation

Risk analysis and objectives

Defining and illustrating an estimation of uncertainty

Risk analysis

The uncertainty of the classification is always given in the output of the AI system. The risks of high uncertainty can be misclassification of a given emotion. The wider system does not use the percentage value, it just uses the prediction with the highest value as the result. This could lead to a degraded user experience. However, the overall risk is low.

Objectives

An uncertainty estimation is implemented in the form that is described before. The exact objectives of the estimation are not clearly defined. This also stems from the fact, that the calculated uncertainty is not used in the wider system.

Criteria for achieving objectives

Documentation of uncertainty metrics and uncertainty estimation quality

There are no metrics assessing the uncertainty estimation. The estimation is taken as is. It can be assumed that the estimation's error is constant. All in all, this topic is insufficiently discussed.

Measures

Data

Choice of a data set annotated with uncertainties

Uncertainties were not part of the training and test data. There are uncertainties in how the labels were assigned. This is briefly discussed by the dataset authors. However, actual uncertainty information was not part of the used dataset. This measure is therefore not usable for this application.

AI component

Selecting an appropriate uncertainty estimation method

The method used for uncertainty estimation is very basic. The values of the output weights used for classification are used. This does not add any information to the information given by the core model itself. It would be insufficient, but this output is not used in the wider system. Therefore, and due to the low risk, this is acceptable.

Post-processing for calibration

There is no post-processing or calibration built into the uncertainty measurement. This is acceptable for the same reasons as before.

Testing the uncertainty estimation

The uncertainty estimation is not tested separately. This is acceptable for the same reasons as before.

Embedding

Assessing follow-up responses

There are no follow-up responses that are implemented or needed in the embedding. Therefore, no documentation is necessary.

Overall assessment

Overall assessment

The documentation and implementation of the uncertainty estimation are very limited. Due to the fact, that the output of the uncertainty estimation is not used and due to the low risk, this is acceptable.

1.5 Control of dynamics

Risk analysis and objectives

Risk analysis

The model is not retrained with new data after its initial setup. Therefore, there is no additional risk in this area. The model will not drift overtime. Concept drift, will also not happen, as human emotions and their facial expressions will not change over time.

Objectives

As there is no additional risk, there is no need to look into this topic further or create objectives.

Criteria for achieving objectives

Intervals and quality requirements for assessing during operation

/

Measures

Measures for operation

Avoiding catastrophic forgetting on new training data

/

Relearning with newly available training data

/

Regular review of the AI application

/

Overall assessment

Overall assessment

As this risk area does not pose any additional risk, it is not necessary to examine it.

1.6 Summary of the dimension

The Reliability dimension of this AI application is generally well-addressed, though with notable gaps in documentation and some areas that could benefit from further exploration. The application domain is clearly defined, and the input data is sufficiently specified, but examples and detailed descriptions are lacking. Risk analysis is adequately performed, particularly in terms of assessing the impact of disturbances, though residual risks and specific objectives are not documented. The quality and origin of the training and test data are well-examined, ensuring that the data is suitable and free from significant biases, with measures taken to prevent data leakage. However, the lack of detailed exploration into corner cases, uncertainty estimation, and real-world generalization testing presents potential concerns. The robustness of the system is sufficiently discussed within the narrow application boundary, though documentation of testing procedures is incomplete. In areas such as intercepting errors at the model level and uncertainty estimation, the risks are deemed low due to the nature of the application, and as such, these areas were not extensively explored. Overall, while the AI system's reliability is considered sufficient for its intended low-risk application, the documentation and some aspects of the analysis could be strengthened to provide a more comprehensive understanding and assurance of reliability. All in all, there are some non-negligible risks, but these risks are acceptable for the application that is examined here.

2 Dimension: Fairness

2.1 Risk area: Fairness

Risk analysis and objectives

Identifying potentially disadvantaged groups

There is only very little documentation available, that identifies potential groups that are disadvantaged in this application. The test and training data that is used has no labels or any information apart from the emotion that is present. Therefore, the ML model cannot use any such information, to unfairly discriminate against certain groups. However, there could be discrimination even without special data. Something that masks facial features, such as hats, beards, hijabs or any other objects of that sort, could potentially impact the quality of the prediction. The colour of a user's skin could feasibly also impact the prediction quality. These examples could be an issue. These potential issues are not addressed in the documentation.

Determining a suitable fairness approach

Fairness, in this application context, means that the emotions of all users are equally well predicted. There is no situation where, for example, a person of colour should have a different prediction outcome for the same a motion as anyone else. Therefore, the prediction quality has to be similar for all groups. The objective therefore is that all potentially affected groups have the same prediction quality. All of this is only insufficiently discussed in the documentation.

Criteria for achieving objectives

Quantifying fairness in the output

There is no documentation defining the groups. There is also no documentation describing metrics or descriptions of how to assess these issues. There are also no target intervals. That makes this point only insufficiently fulfilled.

Quantifying fairness in the training data

There is some documentation available looking into detail on the training data, according to its fairness. As the dataset does not have labels that would make it structurally unfair, this could be sufficient. However, the actual fairness of the training data cannot accurately be judged this way. As there is no data to judge the training data, there are also no target intervals. That makes this point only insufficiently fulfilled.

Measures

Data

Checking data for bias

The data was checked for biases against the emotional labels. There were no checks performed against its bias towards, for example, different skin colours. That makes this point only insufficiently fulfilled.

Fair data pre-processing

There is no preprocessing done, that could affect the fairness of the application. All that is done is image resizing, that should not have any influence in this domain. This makes this topic sufficiently discussed.

AI component

Fair modelling

The fairness as to the above-mentioned groups of users was not separately discussed. The model in general has no inherent structures that would make it structurally unfair. This makes this topic sufficiently discussed.

Fair adaption and post-processing

As there are no target intervals specified, this is not done. This makes this topic sufficiently discussed.

Testing the AI component on unseen data

The AI component was tested with previously unseen data. As there are no specified target intervals, the measurements required cannot be done. This makes this topic sufficiently discussed.

Embedding

Fair further processing

The Embedding does not perform data processing that has the potential to lead to unfair processing. Therefore, this does not need to be looked into.

AI application tests

AI application tests were performed in the real-world scenario. These tests did not show a problem regarding discrimination. However there is no documentation of these tests. This makes this topic sufficiently discussed.

Measures for operation

Monitoring outputs in operation

The output of the operation is not monitored in real time by humans, as this would not be practical for this application. However, due to the relatively low risk of the output, this can be omitted.

Overall assessment

Overall assessment

Overall, not all the required precautions have been met. The entire application is only insufficiently tested for some possible discriminatory situations. This makes this risk area only insufficiently discussed.

2.2 Risk area: Control of dynamics

Risk analysis and objectives

Risk analysis documentation

The AI application does not continue to learn during operation. It is implemented initially and, after setup, left as is. Therefore, there are no particular risks that arise. Concept drift can also be ruled out in this scenario, as human emotions and their corresponding facial features will not change over time. As there is no continuous learning, there are no objectives set. This is sufficient under these circumstances. Therefore, a detailed analysis of this risk area is not necessary.

Criteria for achieving objectives

Maintaining AI application fairness

/

Maintaining fairness in training data

/

Measures

Data

Monitoring training data

/

Measures for operation

Application monitoring

/

Application improvement

/

Monitoring external factors

/

Overall assessment

Overall assessment

This risk area is not relevant for this application, as the ML model does not learn during operation and as concept drift can be ruled out.

2.3 Summary of the dimension

The dimension of Fairness in this AI application highlights significant gaps in addressing potential biases and discrimination risks. Although the model's fairness in predicting emotions across different user groups, such as those with varying skin colours or facial features, is acknowledged as a concern, there is insufficient documentation and analysis to ensure equal prediction quality for all users. The training data lacks necessary labels to evaluate fairness comprehensively, and key metrics or target intervals are not established, leading to an inadequate assessment of fairness. Despite some areas being considered sufficiently discussed, the overall approach to mitigating potential discriminatory outcomes remains insufficiently addressed.

3 Cross-dimensional assessment

Two dimensions were thoroughly examined during this certification process: Fairness and Reliability. No potential trade-offs between these two dimensions were identified, primarily due to the narrow scope of the application. The limitations of the test data also played a significant role. For instance, it was not possible to assess discrimination based on different skin tones. If such information had been available in the datasets, it could have been utilized to test for discrimination and potentially enhance robustness. However, this might have introduced additional challenges in the fairness dimension, which would have necessitated a cross-dimensional assessment. In the current scenario, however, such an assessment is not required.

4 Conclusion

In this test certification, only two dimensions out of the four required by the protection requirement analysis were investigated. This makes a full certification impossible at this point. However, if we only examine the two dimensions, Reliability and Fairness, a conclusion can be drawn. The reliability of the system is generally well-established, with the application operating effectively within its defined scope, the assessment of fairness is found to be insufficient. Significant gaps exist, particularly in addressing potential biases, such as the inability to evaluate discrimination based on skin tones due to the limitations of the test data. These shortcomings indicate that, although the application may be reliable, it does not fully meet the required standards for fairness. As a result, the application cannot be fully certified without further improvements in its approach to fairness.