Towards EU AI Act Compliance: Self-Certification and Fairness Alignment for Facial Emotion Recognition

GREGOR AUTISCHER, Know Center Research GmbH & TU Graz, Austria
KERSTIN WAXNEGGER, Know Center Research GmbH, Austria
DOMINIK KOWALD, Know Center Research GmbH & University of Graz, Austria

The European Union's AI Act establishes comprehensive requirements for high-risk AI systems, yet the harmonized standards granting presumption of conformity remain under development. We investigate the practical application of the Fraunhofer AI Assessment Catalogue as a self-certification framework by conducting a complete self-certification cycle of an AI-based facial emotion recognition system. Starting from a baseline model with deficiencies including inadequate fairness alignment and high prediction uncertainty, we document an enhancement process guided by certification requirements. The enhanced system achieves improved accuracy, higher prediction confidence and comprehensive fairness across demographic groups, successfully satisfying certification criteria for the reliability and fairness dimensions. We find that the certification framework provides substantial value as a proactive development tool, driving concrete technical improvements. However, fundamental gaps separate structured self-certification from legal compliance: harmonized European standards are not yet fully available and assessment catalogues cannot substitute for mandatory conformity assessment procedures. These findings establish the Fraunhofer AI Assessment Catalogue as a valuable preparatory tool that complements rather than replaces formal AI Act compliance requirements.

Keywords: Self-Certification, AI Act, Algorithmic Fairness, AI Reliability, Certification Catalogues

Reference Format:
Gregor Autischer, Kerstin Waxnegger, and Dominik Kowald. 2026. Towards EU AI Act Compliance: Self-Certification and Fairness Alignment for Facial Emotion Recognition. In *Proceedings of Fifth European Conference on Algorithmic Fairness (ECAF'26).* Proceedings of Machine Learning Research, 7 pages.

1 Introduction and Motivation

Artificial intelligence (AI) has moved from research environments into widespread deployment across domains including healthcare, finance, recommender systems and automated decision-making [25, 26]. Virtual assistants can help with daily interactions. Algorithmic systems influence credit and hiring decisions, while AI models optimize energy use [23]. This expansion raises concerns regarding fairness, transparency and accountability, particularly for systems affecting fundamental rights such as labor markets, biometric identification, and emotion recognition [11, 15, 34]. The complexity and opacity of modern AI systems complicate oversight, motivating comprehensive

Authors' Contact Information: Gregor Autischer, Know Center Research GmbH & TU Graz, Graz, Austria, gregor.autischer@student.tugraz.at; Kerstin Waxnegger, Know Center Research GmbH, Graz, Austria, kwaxnegger@know-center.at; Dominik Kowald, Know Center Research GmbH & University of Graz, Graz, Austria, dkowald@know-center.at.

ECAF'26, September 02–September 04, 2026, Ghent, BE

regulatory frameworks [10]. The European Union responded with the AI Act, establishing a risk-based ruleset in which obligations scale with potential harm [19]. Similar regulatory initiatives have emerged globally [4]. Under the AI Act, high-risk AI systems must therefore undergo conformity assessment. However, AI certification differs from traditional software auditing, as adaptive behavior and data dependence require distinct evaluation approaches in a still-developing field with limited practical guidance [12, 40].

This paper summarizes our recent work [9], building on earlier attempts [7, 8] to conduct sample certifications of the RIOT installation's EmoPy facial emotion recognition system using the Fraunhofer AI Assessment Catalogue [32, 33, 38]. Our earlier efforts identified certification gaps but could not resolve them because EmoPy was no longer maintained. In this paper, we rebuild the system in PyTorch as EmoTorch [6], address the identified shortcomings and conduct a complete self-certification. This reconstruction makes the previously unmaintained component modifiable and reproducible, enabling the certification-driven improvements examined in this paper. We pursue three research questions:

RQ1: What practical insights emerge from conducting a complete self-certification cycle using the Fraunhofer AI Assessment Catalogue?

RQ2: What does technical testing reveal that documentation-only certification can miss?

RQ3: To what extent does this self-certification address AI Act requirements regarding reliability and fairness?

2 Background

The AI Act requires conformity assessment for high-risk AI systems prior to market placement, with harmonized standards granting presumption of conformity. However, delays in standardization significantly affect the current certification landscape.

AI Act, Standards and Certification. The AI Act entered into force on August 1, 2024 [19]. It applies a risk-based framework with differentiated requirements depending on potential harm. Facial emotion recognition systems qualify as high-risk under Article 6(2) in conjunction with Annex III, point 1(c). They must comply with Articles 9 to 15 covering risk management, data governance, documentation, transparency, human oversight and accuracy, robustness and cybersecurity [19]. Certification depends on measurable standards against which systems can be assessed [39]. CEN-CENELEC JTC 21 is developing harmonized European standards [1, 2], but delays beyond the original 2025 timeline extend work into 2026 [24]. Until standards are published in the Official Journal, presumption of conformity under Article 40 cannot be used [27, 31]. Existing standards such as ISO/IEC 42001 provide organizational guidance but do not constitute technical specifications for individual AI systems and therefore do not ensure AI Act compliance [17, 21].

Fraunhofer AI Assessment Catalogue. The Fraunhofer AI Assessment Catalogue was developed in anticipation of the AI Act but finalized before its adoption [33]. It offers a structured, questionnaire-based evaluation across six trustworthiness dimensions: fairness; autonomy and control; transparency; reliability; safety and security; and data protection. We use it because it is publicly available, product-oriented and highly granular, with a systematic procedure for assessing AI-specific risks, formulating criteria and metrics, documenting life-cycle measures and reaching an overall assessment. Unlike ISO/IEC 42001, which focuses on organizational management systems, the catalogue provides system-level technical and operational criteria. However, it is not a harmonized standard and does not fulfill mandatory conformity assessment procedures under the AI Act [20].

3 Methodology

We applied the Fraunhofer AI Assessment Catalogue to a facial emotion recognition system through a complete self-certification cycle. The system is the core of the RIOT art installation [37], which uses facial emotion recognition to create a responsive narrative in which audience emotions influence film progression [22]. Participants stand before a screen while a webcam captures facial expressions and the AI adapts the film accordingly [36]. RIOT originally used the EmoPy framework [5]. We reconstructed the original convolutional neural network architecture in PyTorch as EmoTorch [6], establishing a baseline model that replicates the original system while enabling modifications.

Our self-certification focused on reliability and fairness, identified as medium risk in the protection requirement analysis: the system processes facial images and categorizes a person's expression, while misclassification affects the installation experience but does not directly decide access to essential services. The methodology consisted of three phases: (1) baseline assessment using shortcomings identified in the previous certification, (2) systematic enhancements to architecture, training and dataset composition and (3) reassessment of the enhanced model using the same catalogue criteria. This approach integrates certification into the development process, with assessment findings directly informing system improvements. In practice, the catalogue functioned as a feedback loop: initial certification findings were translated into concrete development tasks and implemented through dataset curation, metric reporting, robustness tests and architectural changes. The system was then reassessed against the same criteria. As a self-certification, it differs from independent third-party certification. To ensure reproducibility [35], the full implementation, testing results and documentation are publicly available [6].

4 System Enhancement and Certification Results

In this application, fairness does not mean that predicted emotions should be equally distributed across demographic groups. Rather, users expressing the same target emotion should receive similarly reliable predictions regardless of gender, race, age or visible characteristics such as skin tone and head coverings. We therefore operationalize fairness through group-wise prediction quality: subgroup accuracies, accuracy gaps, disparate-impact style ratios and equalized-odds analysis.

The baseline model, a CNN with two convolutional blocks (four convolution layers) and 6,964 parameters trained on FER-2013 and CK+ [16, 29], was built from 20,940 original images across four emotion categories (anger, fear, calm, surprise) and achieved 53.88% validation accuracy. Technical testing revealed issues masked by accuracy alone: high entropy (1.38), indicating substantial uncertainty, as well as weak subgroup-level performance in the equalized-odds analysis, with TPR values as low as 27.6%. The original datasets lacked demographic metadata, preventing systematic fairness analysis across gender, race and age; subsequent FairFace-based inspection also indicated skew toward Caucasian and middle-aged groups.

The catalogue requirements directly shaped the remediation. Fairness criteria required identifying potentially disadvantaged groups, defining application-specific fairness and quantifying output fairness; reliability criteria required more than aggregate accuracy, including robustness and uncertainty documentation. To address these deficiencies, we expanded the architecture to three convolutional blocks (32, 64, 128 filters) with batch normalization, dropout and Global Average Pooling, increasing parameters to over 300,000. Training improvements included weighted CrossEntropyLoss, AdamW with weight decay, class balancing via WeightedRandomSampler and gradient clipping. We expanded the dataset with RAF-DB [28] to 23,630 original images (259,930 with augmentation), applying

targeted augmentations simulating deployment conditions (dark lighting, blur, noise, head covering). Demographic annotations for gender, race and age were added using the FairFace project.

Here, an accuracy gap denotes the difference between the best- and worst-performing subgroup within a demographic attribute, while the disparate-impact style ratio compares lower- and higher-performing subgroup accuracies. The enhanced model achieves 68.19% test accuracy (macro F1: 0.676), improved confidence (78.7%) and reduced entropy (0.53). Gender gaps are 3.56 percentage points, racial gaps remain within 7.85 percentage points and performance exceeds 68% for age groups 0 to 69, with lower performance for 70+ (45.61%). Augmentation testing confirms robustness across conditions, with weaker performance only for heavily blurred images (54.89%).

Reassessment using the Fraunhofer Catalogue indicates that the enhanced system meets certification requirements for reliability and fairness. Table 1 summarizes how selected catalogue focus areas were translated into technical changes and evidence.

Table 1. From catalogue criteria to system enhancements and certification evidence.

Catalogue focus area	**Baseline gap**	**Remediation**	**Evidence**
Fairness: identify potentially affected groups and define fairness	Original datasets lacked demographic metadata; fairness was not operationalized	Added FairFace demographic annotations; defined fairness as comparable prediction quality	Race gap 7.85%, gender gap 3.56%, disparate-impact ratio 0.888
Fairness: quantify output fairness on unseen data	Original documentation lacked demographic fairness metrics; baseline subgroup performance was weak	Added per-group accuracy, confusion matrices and age/race/gender test-set analysis	All race groups above 60%; age groups 0 to 69 above 68%; 70+ at 45.61%
Reliability: standard-case performance and uncertainty	Aggregate accuracy documentation hid low confidence and high entropy	Added macro F1, confidence and entropy; redesigned model and training	68.19% accuracy, 0.676 macro F1, 78.7% confidence, 0.53 entropy
Reliability: robustness and application boundary	Baseline documentation reported no augmentation or robustness analysis	Added deployment-style tests for lighting, contrast, noise, blur and head coverings	All augmentations above 54%; weakest case blur at 54.89%

5 Key Findings and Compliance Gap

RQ1: Practical insights from complete self-certification. The Fraunhofer Catalogue's structured criteria required deeper analysis than accuracy documentation alone. Accuracy metrics obscured low prediction confidence, while fairness assessment required demographic metadata absent from the original datasets. For legacy systems, this matters because systems that cannot be modified must meet requirements as they are, as demonstrated by the baseline EmoPy system. Beyond evaluation, the catalogue served as a development framework: subgroup labels and metrics followed from fairness criteria, while entropy, confidence and augmentation tests followed from reliability criteria. Integrating certification requirements during development also produced structured documentation as a byproduct, strengthening assessment validity.

RQ2: Technical testing versus documentation-only certification. Technical implementation revealed issues that documentation alone did not detect. The baseline model showed acceptable accuracy while masking high entropy and low confidence, with approximately 27% confidence for a four-class prediction problem. Accuracy-focused documentation overlooked these characteristics. Enhanced evaluation using entropy, confidence, demographic fairness and robustness metrics provided a more comprehensive picture of system behavior. Deeper technical analysis therefore complements the documentation-based Fraunhofer Catalogue and reduces oversight risk.

RQ3: Relationship to AI Act requirements. Significant gaps prevent using the Fraunhofer Catalogue as a standalone means of demonstrating compliance with the AI Act. Harmonized European standards remain under development by CEN-CENELEC JTC 21, with delays beyond the original 2025 target [3]. The Fraunhofer Catalogue alone is insufficient to demonstrate conformity with the AI Act. [20]. AI Act conformity assessment requires procedures leading to the EU Declaration of Conformity [18, 19], Article 17 quality management systems, post-market monitoring, and the applicable conformity assessment before the system is placed on the market once the relevant AI Act obligations apply. Despite these gaps, structured self-certification has preparatory value by aligning improvements and documentation with AI Act principles and facilitating future transition to formal conformity assessment once harmonized standards are available.

6 Conclusion and Future Research Directions

Our complete self-certification cycle using the Fraunhofer AI Assessment Catalogue demonstrates its effectiveness as both an evaluation tool and a development framework for a facial emotion recognition system. The enhanced system satisfies certification requirements for reliability and fairness, with improvements directly driven by catalogue criteria: architectural refinement reduced prediction uncertainty, expanded datasets enabled demographic fairness analysis and systematic documentation emerged throughout development.

However, substantial gaps remain between structured self-certification and AI Act compliance. The absence of harmonized European standards prevents presumption of conformity. Furthermore, the Fraunhofer Catalogue is not a recognized conformity assessment mechanism under the AI Act and cannot replace the applicable conformity assessment procedures. This highlights the distinction between internal quality assurance and formal legal compliance. Even after harmonized standards become available, structured self-certification remains useful as an internal development and documentation practice, provided it is clearly separated from formal legal compliance.

Future work could examine how certification frameworks evolve once harmonized standards become available, explore transitions from self-certification to formal conformity assessment and assess commercially deployed AI systems to better understand real-world compliance challenges (e.g., with respect to multiple stakeholders that might be impacted by the AI decisions [13, 14, 30]).

Acknowledgments

Know Center is a COMET competence center funded by the Austrian Federal Ministry of Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK), the Austrian Federal Ministry of Labour and Economy (BMAW), the State of Styria, SFG, the Vienna Business Agency and Standortagentur Tirol. The COMET programme is managed by the Austrian Research Promotion Agency (FFG).

References

[1] CEN-CENELEC JTC 21. 2024. European AI Standardization | CEN-CENELEC JTC 21 |. https://jtc21.eu/

[2] CEN-CENELEC JTC 21. 2025. Artificial Intelligence. https://www.cencenelec.eu/areas-of-work/cen-cenelec-topics/artificial-intelligence/

[3] EU Artificial Intelligence Act. 2025. Standard Setting | EU Artificial Intelligence Act. https://artificialintelligenceact.eu/standard-setting-overview/

[4] Amir Al-Maamari. 2025. Between Innovation and Oversight: A Cross-Regional Study of AI Risk Management Frameworks in the EU, U.S., UK, and China. doi:10.48550/arXiv.2503.05773 arXiv:2503.05773 [cs].

[5] Angelica Perez, Julien Deswaef, and Puneetha Pai. 2021. thoughtworksarts/EmoPy. https://github.com/thoughtworksarts/EmoPy original-date: 2017-12-20T02:19:22Z.

[6] Gregor Autischer. 2025. gregor-autischer/EmoTorch_Paper. https://github.com/gregor-autischer/EmoTorch_Paper original-date: 2025-09-11T09:35:41Z.

[7] Gregor Autischer, Kerstin Waxnegger, and Dominik Kowald. 2025. AI Certification and Assessment Catalogues: Practical Use and Challenges in the Context of the European AI Act. In *Proceedings of EWAF*. 492–498. https://proceedings.mlr.press/v294/autischer25a.html

[8] Gregor Autischer, Kerstin Waxnegger, and Dominik Kowald. 2025. Practical Application and Limitations of AI Certification Catalogues in the Light of the AI Act. doi:10.48550/arXiv.2502.10398 arXiv:2502.10398 [cs].

[9] Gregor Autischer, Kerstin Waxnegger, and Dominik Kowald. 2026. Self-Certification of High-Risk AI Systems: The Example of AI-based Facial Emotion Recognition. doi:10.48550/arXiv.2601.08295 arXiv:2601.08295 [cs].

[10] Tahir Ayub. 2024. Regulating AI: Balancing Innovation, Ethics, and Public Policy A Critical Analysis of AI Governance and Policy Approaches. doi:10.2139/ssrn.5190225

[11] Kevin Baum, Joanna Bryson, Frank Dignum, Virginia Dignum, Marko Grobelnik, Holger Hoos, Morten Irgens, Paul Lukowicz, Catelijne Muller, Francesca Rossi, John Shawe-Taylor, Andreas Theodorou, and Ricardo Vinuesa. 2023. From fear to action: AI governance and opportunities for all. *Frontiers in Computer Science* 5 (May 2023). doi:10.3389/fcomp.2023.1210421 Publisher: Frontiers.

[12] Kiera Brogle, Eva Kallina, Haley Sargeant, et al. 2025. Context-specific certification of AI systems: a pilot in the financial industry. *AI and Ethics* 5 (2025), 4223–4240. doi:10.1007/s43681-025-00720-w

[13] Robin Burke, Gediminas Adomavicius, Toine Bogers, Tomasso Di Noia, Dominik Kowald, Julia Neidhardt, Özlem Özgöbek, Maria Pera, and Jürgen Ziegler. 2024. Multistakeholder and multimethod evaluation. In *Evaluation Perspectives of Recommender Systems: Driving Research and Education (Dagstuhl Seminar 24211)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik GmbH, 123–145.

[14] Robin Burke, Gediminas Adomavicius, Toine Bogers, Tommaso Di Noia, Dominik Kowald, Julia Neidhardt, Özlem Özgöbek, Maria Soledad Pera, Nava Tintarev, and Jürgen Ziegler. 2025. De-centering the (Traditional) user: Multistakeholder evaluation of recommender systems. *International Journal of Human-Computer Studies* 203 (2025), 103560. doi:10.1016/j.ijhcs.2025.103560

[15] Ben Chester Cheong. 2024. Transparency and accountability in AI systems: safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics* 6 (July 2024). doi:10.3389/fhumd.2024.1421273 Publisher: Frontiers.

[16] Jeffrey Cohn. 2024. Resources – Jeffrey Cohn. https://www.jeffcohn.net/resources/

[17] Yalcin Erleblebici. 2025. ISO 42001 vs. AI Act. https://www.activemind.de/magazin/iso-42001-ai-act/

[18] EU AI Act. 2024. Key Issue 2: Conformity Assessment and Self-Assessment - EU AI Act. https://www.euaiact.com/key-issue/2

[19] European-Union. 2024. 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. http://data.europa.eu/eli/reg/2024/1689/oj/eng Legislative Body: CONSIL, EP.

[20] Future of Privacy Forum. 2025. Conformity Assessments under the EU AI Act: A step-by step guide. https://fpf.org/wp-content/uploads/2025/04/OT-comformity-assessment-under-the-eu-ai-act-WP-1.pdf

[21] ISO. 2023. ISO/IEC 42001:2023. https://www.iso.org/standard/42001

[22] Karen Palmer. 2017. RIOT Video. https://www.youtube.com/watch?v=-BCny9SuI3A

[23] Maria Kasinidou, Styliani Kleanthous, Matteo Busso, Marcelo Rodas, Jahna Otterbacher, and Fausto Giunchiglia. 2024. Artificial Intelligence in Everyday Life 2.0: Educating University Students from Different Majors. In *Proceedings of ITiCSE (ITiCSE 2024)*. New York, NY, USA, 24–30. doi:10.1145/3649217.3653542

[24] Robert Kilian, Linda Jäck, and Dominik Ebel. 2025. European AI Standards – Technical Standardisation and Implementation Challenges under the EU AI Act. *European Journal of Risk Regulation* 16, 3 (Sept. 2025), 1038–1062. doi:10.1017/err.2025.10032 Publisher: Cambridge University Press.

[25] Florian Königstorfer, Armin Haberl, Dominik Kowald, Tony Ross-Hellauer, and Stefan Thalmann. 2024. Black Box or Open Science? A Study On Reproducibility In AI Development Papers. In *57th Annual Hawaii International Conference on System Sciences, HICSS 2024*. IEEE Computer Society, 682–691.

[26] Dominik Kowald, Sebastian Scher, Viktoria Pammer-Schindler, Peter Müllner, Kerstin Waxnegger, Lea Demelius, Angela Fessl, Maximilian Toller, Inti Gabriel Mendoza Estrada, Ilija Šimić, et al. 2024. Establishing and evaluating trustworthy AI: overview and research challenges. *Frontiers in Big Data* 7 (2024), 1467222.

[27] Andrew Leyden. 2025. Standards and the EU AI act: legitimacy, state of play, and future challenges. *Information & Communications Technology Law* 0, 0 (2025), 1–31. doi:10.1080/13600834.2025.2570966 Publisher: Routledge.

[28] Shan Li and Weihong Deng. 2019. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition. *IEEE Transactions on Image Processing* 28, 1 (2019), 356–370.

[29] Microsoft. 2023. microsoft/FERPlus. https://github.com/microsoft/FERPlus original-date: 2016-09-14T06:35:21Z.

[30] Peter Müllner, Anna Schreuer, Simone Kopeinik, Bernhard Wieser, and Dominik Kowald. 2025. Multistakeholder fairness in tourism: what can algorithms learn from tourism management? *Frontiers in Big Data* Volume 8 - 2025 (2025). doi:10.3389/fdata.2025.1632766

[31] Claire O'Brien, Mark Rasdale, and Daisy Wong. 2024. The role of harmonised standards as tools for AI act compliance. https://www.dlapiper.com/en/insights/publications/2024/01/the-role-of-harmonised-standards-as-tools-for-ai-act-compliance

[32] Angelica Perez. 2018. EmoPy: a machine learning toolkit for emotional expression. https://www.thoughtworks.com/insights/blog/emopy-machine-learning-toolkit-emotional-expression

[33] Dr Maximilian Poretschkin, Anna Schmitz, Dr Maram Akila, Linara Adilova, Dr Daniel Becker, Dr Armin B Cremers, Dr Dirk Hecker, Dr Sebastian Houben, Julia Rosenzweig, Joachim Sicking, Elena Schulz, Dr Angelika Voss, and Dr Stefan Wrobel. 2023. *AI Assessment Catalog*. Technical Report. Fraunhofer IAIS.

[34] Sebastian Scher, Simone Kopeinik, Andreas Trügler, and Dominik Kowald. 2023. Modelling the long-term fairness dynamics of data-driven targeted help on job seekers. *Scientific Reports* 13, 1 (2023), 1727.

[35] Harald Semmelrock, Tony Ross-Hellauer, Simone Kopeinik, Dieter Theiler, Armin Haberl, Stefan Thalmann, and Dominik Kowald. 2025. Reproducibility in machine-learning-based research: Overview, barriers, and drivers. *AI Magazine* 46, 2 (2025), e70002.

[36] TED Residency. 2018. Karen Palmer: The film that watches you back. https://www.youtube.com/watch?v=Rw8gLEkFdSw

[37] Thoughtworks. 2018. RIOT | Thoughtworks Arts. https://thoughtworksarts.io/projects/riot/

[38] Thoughtworks. 2019. thoughtworksarts/riot. https://github.com/thoughtworksarts/riot original-date: 2017-11-22T15:59:17Z.

[39] Susanne Werry, Simon Ridgway, Mark Kerr-Shaw, and Andrew Silverstein. 2024. EU Standardization Supporting the Artificial Intelligence Act | Insights | Skadden, Arps, Slate, Meagher & Flom LLP. https://www.skadden.com/insights/publications/2024/10/eu-standardization-supporting-the-artificial-intelligence-act

[40] Philip Matthias Winter, Sebastian Eder, Johannes Weissenböck, Christoph Schwald, Thomas Doms, Tom Vogt, Sepp Hochreiter, and Bernhard Nessler. 2021. Trusted Artificial Intelligence: Towards Certification of Machine Learning Applications. doi:10.48550/arXiv.2103.16910 arXiv:2103.16910 [cs, stat].