Journal of Law, Market & Innovation Vol. 4 - Issue 3/2025

Chiara Gallese[*], *Elena Falletti*[*]

GENERAL SECTION

DATASET ASSESSMENT TOOLS FOR THE AI ACT COMPLIANCE OF HIGH-RISK AI SYSTEMS[*]

Abstract

The European Union's Artificial Intelligence Act (AI Act), effective as of August 2024, introduces stringent data governance requirements for high-risk AI systems. Central to these obligations is the mandate for data used in training, validation, and testing to be subject to rigorous quality controls and governance practices aligned with the system's intended purpose. This paper examines the AI Act's novel emphasis on responsible data handling as a foundation for mitigating bias, ensuring safety, and protecting fundamental rights. It argues that while existing scholarship largely focuses on algorithmic fairness at the model level, insufficient attention has been given to fairness embedded within the data itself. Addressing this gap, the paper explores the complex ethical and epistemological tensions inherent in defining "fairness", and critiques current policy frameworks for their narrow focus on technical fixes. The study highlights the FAnFAIR tool as a pioneering approach to data fairness evaluation, offering a hybrid methodology that integrates automated statistical diagnostics with qualitative, context-aware assessments. FAnFAIR aligns its features with the AI Act's legal mandates and facilitates compliance through continuous bias monitoring, data quality enhancement, and informed decision-making about dataset suitability. This paper provides a socio-legal and technical analysis of data governance under the AI Act, advocating for a shift toward embedded ethical oversight throughout the AI lifecycle and, in particular, in the data processing phase.

JEL CLASSIFICATION: K10, K13

SUMMARY

[*] Tilburg Institute for Law, Technology, and Society, The Netherlands.
[*] Università Carlo Cattaneo-LIUC, Italy.
[*] Elena Falletti wrote the State of the Art section, Chiara Gallese wrote the rest. All authors reviewed the draft.

10.13135/2785-7867/12840

1 Introduction

Entered into force in August 2024, the Artificial Intelligence Act (AI Act)[1] prescribes several requirements and obligations for high-risk systems, particularly concerning Data Governance. Such systems must adhere to specific quality criteria related to training, validation, and testing, which must be “subject to data governance and management practices appropriate for the intended purpose of the high-risk AI system” under Article 10(2) AI Act. These requirements aim to ensure responsible data handling and mitigate risks related to bias, safety, and security issues that are known to affect AI systems.

Firstly, data governance must reflect the design choices made during the creation of the AI system, aligning data processing practices with the intended purpose of the system, such as considering which data sources and processing methods are most appropriate for achieving the system’s goals. The processes for data collection must be transparent and well-documented, for example regarding the origin of the data.

Data governance also involves a continuous oversight of data preparation activities. This includes operations such as annotation, labelling, cleaning, updating, enrichment, and aggregation of data. These processes must ensure that the data is of high quality and suitably transformed for the AI system's requirements. Additionally, there must be a clear understanding of the assumptions made about what the data is intended to measure or represent, with an evaluation of whether it accurately reflects the real-world phenomena it aims to model.

Another essential aspect of the data governance framework is assessing the availability, quantity, and suitability of the data. This assessment helps to verify that the data used is complete and reliable, minimising the risk of errors and deficiencies. The AI Act places great emphasis on identifying potential biases within the data that could affect the health and safety of individuals, infringe on fundamental rights, or lead to discrimination prohibited under Union law, deeming that addressing these biases is important to mitigating unfair or harmful outcomes.

In instances where biases are detected, the AI Act prescribes that appropriate measures must be taken to prevent and mitigate their effects. This could involve revising the data collection processes, enhancing the quality of the data, or implementing fairness checks throughout the AI system’s development. Moreover, it is vital to identify any gaps or shortcomings in the data that might hinder compliance with the AI Act. Once these issues are acknowledged, strategies can be developed to address them, ensuring that the AI system maintains a high level of quality.

The predominant focus in the academic literature has been on uncovering biases within the models themselves and biases in the use of AI systems, as opposed to biases inherent

[1] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

in the data. This focus has resulted in creating a wide array of metrics designed to assess the fairness of algorithmic results. Assessing algorithmic fairness within the data context was relatively limited, primarily focusing on transparency and documentation procedures.

The examination of algorithmic fairness within datasets has followed two primary research directions. One strand of research is concerned with scrutinizing data curation processes, and the other one with the empirical analysis of fairness-related data aspects, such as how data can impact the fairness of algorithms.

The case law about discrimination in datasets is scarce since parties tend to settle out of court to avoid negative publicity. This has been the case in the United States, as seen in the agreement approved by the Southern District Court of New York between Meta and the Department of Justice concerning discrimination in social housing advertising, which violated the Fair Housing Act[2]. In Europe, a notable example of tools for detecting algorithmic bias is the Dutch experience with the Algorithm Audit conducted by DUO (Dienst Uitvoering Onderwijs)[3]. This audit was used to review scholarship recipient selection procedures to identify racial bias against applicants with a migration background.

As highlighted by the European Digital Rights (EDRi) report, current policy frameworks often adopt a constrained understanding of fairness in AI, focusing narrowly on the technical remediation of datasets, such as eliminating bias or ensuring that data is “representative, error-free, and complete”, without addressing the broader societal structures in which AI systems operate.[4] This approach, while necessary, is not sufficient. AI systems trained on “debiased” datasets may still replicate and legitimise existing inequalities embedded in the institutions, processes, and ideologies that shape the data in the first place. Thus, a technically fair dataset may nonetheless contribute to inequitable outcomes.

This limitation points to a deeper epistemological and ethical tension in the field of algorithmic fairness. As Hanna et al. argue, the discourse must shift from an exclusive focus on the algorithmic layer toward a systemic analysis of the social, political, and institutional contexts in which AI is designed, deployed, and governed.[5] Indeed, while the literature acknowledges the critical role of data quality, representativeness, and statistical soundness in shaping model performance,[6] it simultaneously calls for an

[2] Agreement between Meta and the Department of Justice <https://about.fb.com/wp-content/uploads/2022/06/June_2022_HUD_Settlement_Agreement.pdf> accessed 16 September 2025.

[3] ‘Profiled Without Protection - Students In The Netherlands Hit By Discriminatory Fraud Detection System’ Research Briefing (Amnesty International 2024).

[4] Agathe Balayn and Seda Gürses, ‘Beyond Debiasing: Regulating AI and Its Inequalities’ EDRi Report (EDRi 2021).

[5] Alex Hanna and others, ‘Towards a Critical Race Methodology in Algorithmic Fairness’ in *FAT* ‘20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (ACM 2020).

[6] Venkat N Gudivada, Amy Apon and Junhua Ding, ‘Data Quality Considerations for Big Data and Machine Learning: Going beyond Data Cleaning and Transformations’ (2017) 10 International Journal on Advances in Software 1; R Stuart Geiger and others, ‘Garbage in, Garbage out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes from?’ in *FAT* ‘20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (ACM 2020); Divy Thakkar and others, ‘When Is Machine Learning Data Good?: Valuing

expanded view of fairness that incorporates normative values and interrogates structural power imbalances.

In practice, this means recognising that unfairness is not intrinsic to data alone but is also a product of the human and institutional decisions surrounding data collection, annotation, and reuse. For instance, ensuring informed consent for data reuse,[7] enforcing transparency in secondary data usage,[8] and conducting extensive ethics assessments that go beyond procedural review boards[9] are all critical measures. These actions must be grounded in foundational ethical principles[10] such as accountability, dignity and self-determination, traceability, stakeholder engagement, risk assessment, and attention to societal impact. Embedding such principles across the AI lifecycle[11], from data design to system deployment, is vital for mitigating harm and advancing a truly equitable AI ecosystem.

In the literature, only few tools have been developed to assess datasets for fairness, among which is the FAnFAIR software.[12] Within its framework, the insights drawn from the aforementioned literature are operationalised through a hybrid methodology that combines autonomously computed statistical properties of datasets with qualitative evaluations manually entered by the user. By integrating both quantitative and normative considerations, FAnFAIR enables a context-aware assessment of data fairness. The tool can inform key decisions during the early stages of AI development, including whether to apply pre-processing techniques or, in cases of significant bias or legal non-compliance, whether to discard the dataset entirely to avoid AI-driven harm. In this way, FAnFAIR serves as a foundational resource for guiding principled data governance and risk mitigation, contributing to the creation of lawful, ethical, and socially responsible AI systems.

FAnFAIR offers a new toolset that aligns with the data governance requirements of the AI Act, which can be important especially in the context of medical applications. By

in Public Health Datafication' in *CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (ACM 2022) <https://doi.org/10.1145/3491102.3501868> accessed 15 September 2025.

[7] Chiara Gallese and others, 'Predicting and Characterizing Legal Claims of Hospitals with Computational Intelligence: The Legal and Ethical Implications' in *2022 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* (IEEE 2022) <https://doi.org/10.1109/CIBCB55180.2022.9863033> accessed 15 October 2025.

[8] Chiara Gallese, 'Legal Aspects of AI in the Biomedical Field. The Role of Interpretable Models' in Bruno Carpentieri and Paola Lecca (eds), *Big Data Analysis and Artificial Intelligence for Medical Sciences* (Wiley 2024).

[9] Emmanuel R Goffi, Louis Colin and Saida Belouali, 'Ethical Assessment of AI Cannot Ignore Cultural Pluralism: A Call for Broader Perspective on AI Ethic' (2021) 1 Arribat-International Journal of Human Rights 151.

[10] S Lo Piano, 'Ethical Principles in Machine Learning and Artificial Intelligence: Cases from the Field and Possible Ways Forward' (2020) 7(1) Humanities and Social Sciences Communications 1; B Giovanola and S Tiribelli, 'Beyond Bias and Discrimination: Redefining the AI Ethics Principle of Fairness in Healthcare Machine-Learning Algorithms' (2023) 38(2) AI & Society 549.

[11] I Dankwa-Mullan and others, 'A Proposed Framework on Integrating Health Equity and Racial Justice into the Artificial Intelligence Development Lifecycle' (2021) 32(2) Journal of Health Care for the Poor and Underserved 300.

[12] Chiara Gallese and others, 'Investigating Semi-Automatic Assessment of Data Sets Fairness by Means of Fuzzy Logic' in *2023 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* (IEEE 2023) <https://doi.org/10.1109/CIBCB56990.2023.10264913> accessed 15 September 2025.

providing a semi-automatic assessment of dataset fairness, FAnFAIR addresses critical aspects of data quality, bias detection, and pre-processing decisions, all of which are central to regulatory compliance. FAnFAIR's primary utility lies in its rule-based fuzzy inference system, which combines automated statistical analysis with user-provided qualitative insights. This hybrid approach allows the software to evaluate multiple fairness metrics autonomously, reducing the risk of human error and subjective bias. The AI Act mandates a careful examination of data for biases that might impact health, safety, and fundamental rights, and FAnFAIR's capacity to detect issues related to balance, outliers, missing values, and overall dataset quality helps users identify and address potential problems early in the AI development process, thereby mitigating risks before they propagate through the system.

The software's ability to assist in deciding whether to pre-process a dataset, or even discard it entirely, is aligned with the AI Act's focus on data suitability and representativeness. It also allows users to reflect on their organizational measures and compliance procedures: since it integrates fairness considerations into dataset evaluation, it supports compliance with requirements related to the formulation of data assumptions and the detection of data gaps or shortcomings. This ensures that only datasets that meet stringent fairness criteria proceed to the training stage, mitigating the potential for biased or harmful outcomes.

Recent enhancements to FAnFAIR,[13] including simplified data import, improved outlier detection, and analysis of sensitive variables, further align the software with the AI Act's requirements for thorough data examination. These new features allow for a deeper analysis of factors that could contribute to bias, especially in complex datasets with missing or noisy data. For example, FAnFAIR's testing on real-world medical datasets, such as those involving COVID patients, demonstrates its practical application in identifying how pre-processing steps can influence the fairness and predictive performance of AI models.[14] This capability directly supports the AI Act's obligation for continuous assessment and mitigation of biases throughout the AI system lifecycle.

In the case of the COVID dataset, the initial version included 951 patients and 129 fields:

- health history collected at hospital admission (24), such as age, sex, body-mass index, and 21 risk factors (e.g., smoking habits).
- administered therapies (20)- explaining which therapies were administered to the patient, such as drugs (e.g., warfarin), and different types of ventilatory support (e.g., invasive mechanical ventilation).
- complications (21) – identifying complications that occurred to the patient (e.g., acute renal failure).

[13] Michele Rispoli and others, 'Investigating Fairness with FAnFAIR: Is Pre-Processing Useful Only for Performances?' in *2025 IEEE Symposium on Computational Intelligence in Health and Medicine (CIHM)* (IEEE 2025) <https://doi.org/10.1109/CIHM64979.2025.10969477> accessed 15 September 2025.

[14] The detailed methodology is described in Rispoli and others (n 14).

- serial measurements (45) – detailing up to five measurements for each of nine variables, such as physiological quantities (e.g., lymphocytes), and indicators of the status of pulmonary functions (e.g., arterial partial pressure of oxygen to fraction of inspired oxygen ratio).
- dates (18) – detailing variables such as dates of birth, hospitalisation, decease/discharge, sampling of serial values.
- outcome (1) - annotating patient's death.

The ML task addressed in our article is intended to solve a binary classification problem: the resulting algorithm predicts mortality of each patient, based on the data collected during their hospitalisation. We used FanFAIR to evaluate four versions of our dataset, each in a different stage of the data processing pipeline adopted in the original study[15] we referenced:

1) Raw Dataset - (947 rows, 129 columns) in csv, minimal processing was applied, that is datatype enforcement and automated removal of invalid entries of numeric variables.
2) Cleaned Dataset – (825 rows, 79 columns), in which uninformative and highly sparse variables were removed, along with rows corresponding to patients outside the target population—specifically, those not treated with glucocorticoids (GCs) or presenting clearly anomalous values. Of the nine available serially measured variables, only two were retained: the PaO_2/FiO_2 ratio and C-reactive protein (CRP) levels.
3) Re-serialised Dataset – (825 rows, 57 columns), in which date information was removed, and raw serial measurements were transformed into three summary features per series: the first value, the last value, and a binary "improving" indicator. This indicator was derived from the raw measurements using heuristics developed by medical experts.
4) Selected Dataset – (825 rows, 10 columns), in which only the nine predictor variables identified during the variable selection phase of the original study's processing pipeline were used, along with the outcome variable. Additionally, the inclusion of legal compliance as a metric in FAnFAIR helps in supporting adherence to the regulatory framework set forth by the AI Act. By evaluating whether dataset creators have respected the related laws and duties, the software aids in ensuring that data governance practices align with the applicable laws and regulations. This integrated approach reduces the likelihood of non-compliance and enhances the overall accountability and transparency of AI system development.

FAnFAIR thus provides a robust solution for meeting the AI Act's data governance requirements by streamlining the process of fairness evaluation, offering actionable insights, and supporting informed decision-making about dataset suitability. Its combination of automated analysis and human expertise facilitates a nuanced compliance-oriented approach to data management in high-risk AI systems.

[15] F Salton and others, 'A Tailored Machine Learning Approach for Mortality Prediction in Severe COVID-19 Treated with Glucocorticoids' (2024) 28(9) Int J Tuberc Lung Dis 439.

This article is divided as follows: Section 1 presents our work and its context; Section 2 provides a summary of the state of the art related to dataset assessment tools; Section 3 explains why assessment tools are needed; Section 4 describes the FAnFAIR assessment tool; and Section 5 draws some concluding remarks.

2 State of the art

This article focuses on the data assessment[16] instruments developed or studied since the passing of the AI Act in the EU. We start with an analysis of the articles that have been published since the definitive approval of the AI Act by the European Parliament. Scholarship that has since been confronted with the development of data assessment tools can be divided into two groups: on the one hand, some scholars have engaged in theoretical reconstruction in connection with the provisions of the AI Act relating to AI Systems and datasets; on the other hand, some researchers have developed operational tools, which will be briefly reviewed.

On the theoretical side, there are contributions such as that made by Golpayegani and others,[17] which argue that with the enactment of the AI Act, the documentation of high-risk AI systems and related risk management information will become a legal requirement that will play a key role in demonstrating compliance. Despite this, there is a lack of standards and guidelines to help draft AI and risk documentation in line with the AI Act. The authors propose AICat - an extension of DCAT - for representing catalogues of AI systems that provides consistency, machine-readability, searchability, and interoperability in managing open metadata regarding AI systems.

Butt and others[18] refer to a multidisciplinary approach to examine the complicated provisions and implications of the AI Act, which encompasses legal, ethical, socio-economic, and technological dimensions. An important aspect of this research is how the Regulation addresses the identification and mitigation of biases within AI systems.

Hernandez and others[19] propose a terminological mapping of the AI Act to assess the consistency and clarity of the definitions and concepts contained therein, with a view to facilitating compliance and regulatory harmonisation. Indeed, gaps in definitions can affect the consistency of compliance statements across organisations, sectors and applications. This can lead to regulatory uncertainty, particularly for SMEs and public sector bodies that rely on compliance with standards rather than equivalent proprietary

[16] In the context of this article, “data assessment” refers to the rules, tools, or methods that decide which data goes where and how it is used, if it complies with relevant law or ethical principles, if it is biased or otherwise flawed.

[17] Delaram Golpayegani, Harshvardhan J Pandit and Dave Lewis, ‘AICat: An AI Cataloguing Approach to Support the EU AI Act’ [2024] arXiv:2501.04014 <https://doi.org/10.48550/arXiv.2501.04014> accessed 11 September 2025.

[18] Junaid Sattar Butt, ‘Analytical Study of the World’s First EU Artificial Intelligence (AI) Act’ (2024) 5(3) International Journal of Research Publication and Reviews 7343.

[19] Julio Hernandez, Delaram Golpayegani and Dave Lewis, ‘An Open Knowledge Graph-Based Approach for Mapping Concepts and Requirements Between the EU AI Act and International Standards’ [2024] arXiv:2408.11925 <https://doi.org/10.48550/arXiv.2408.11925> accessed 11 September 2025.

systems for the development and implementation of compliant high-risk AI systems. The paper offers a simple and repeatable mechanism for mapping terms and requirements related to regulatory statements contained in regulations and standards, such as the AI Act and ISO management system standards, to open knowledge graphs to identify where regulatory steps need to be taken from a technical perspective to promote transparency and explainability of regulatory processes.

Davvetas and others[20] elaborate the TAI Scan Tool, that is a RAG-based self-assessment tool that supports the legal TAI assessment, with a particular emphasis on facilitating compliance with the AI Act. It involves a two-step approach with a pre-screening and an assessment phase. The assessment output of the system includes insight regarding the risk-level of the AI system according to the AI Act, while at the same time retrieving relevant articles to aid with compliance and notify on its obligations.

Ceravolo and others[21] propose a Fundamental Rights Impact Assessment (FRIA) methodology, which aims to assess the human rights impact on high-risk AI systems and ensure compliance with the EU AI Act.

The second group of scholars focused on the practical development of data assignment tools. Focusing on logistics, Marino and others[22] introduce ‘Compliance Cards’, a system that automates AI compliance analysis by capturing metadata related to AI systems and their datasets. The metadata is then evaluated using an automated algorithm to predict compliance status. This approach enhances the efficiency of regulatory assessments by providing a structured and automated compliance evaluation process. Golpayegani and others[23] show a similar approach with “AI Cards”, a tool that aligns with the AI Act’s documentation requirements, improving transparency and accountability in AI compliance assessment.

Milaj and others[24] elaborate an open-source compliance tool to be used in migration, asylum and border control management. This study analyses from a legal doctrinal approach the compliance of open-source data from social media platforms with high-quality data and transparency requirements. The authors underline that, since transparency marks the boundary between high-risk and unacceptable-risk AI, using open social media data for border control risk assessment may constitute an unacceptable threat to fundamental rights without proper safeguards.

[20] Athanasios Davvetas and others, ‘TAI Scan Tool: A RAG-Based Tool With Minimalistic Input for Trustworthy AI Self-Assessment’ [2025] arXiv:2507.17514 <https://arxiv.org/abs/2507.17514> accessed 9 September 2025.

[21] Paolo Ceravolo and others, ‘HH4AI: A Methodological Framework for AI Human Rights Impact Assessment under the EUAI Act’ [2025] arXiv:2503.18994 <https://arxiv.org/abs/2503.18994> accessed 9 september 2025.

[22] Bill Marino and others, ‘Compliance Cards: Automated EU AI Act Compliance Analyses Amidst a Complex AI Supply Chain’ [2024] arXiv:2406.14758 <https://doi.org/10.48550/arXiv.2406.14758> accessed 11 September 2025.

[23] Delaram Golpayegani and others, ‘AI Cards: Towards an Applied Framework for Machine-Readable AI and Risk Documentation Inspired by the EU AI Act’ in Meiko Jensen, Cédric Lauradoux and Kai Rannenberg (eds), *Privacy Technologies and Policy: 12th Annual Privacy Forum, APF 2024, Karlstad, Sweden, September 4-5, 2024, Proceedings* (Springer 2024) 2.

[24] Jonida Milaj and other, ‘Transparency as the Defining Feature for Developing Risk Assessment AI Technology for Border Control’ (2025) 39 International Review of Law, Computers & Technology 140.

Bogucka and others[25] develop a template for compliance assessment co-designed with AI practitioners and regulatory experts. This template integrates requirements from the AI Act, NIST's AI Risk Management Framework,[26] and ISO 42001,[27] guiding organisations in evaluating dataset quality and bias to ensure regulatory adherence.

Kelly and others[28] propose a methodological approach to AI Act compliance by extending traditional product quality models to encompass AI-specific regulatory requirements. Their approach systematically maps compliance attributes to quality standards, thereby assisting in structured dataset evaluations for high-risk AI applications.

Chard and others[29] studied an audit framework specifically designed for assessing privacy compliance in AI systems. Their framework includes precise prompts for evaluating privacy practices, enhancing AI technologies' security and trustworthiness. This contribution develops robust data assessment tools that align with regulatory requirements.

There is a relevant area of research in this field that goes beyond academic scholarship, since AI assessment tools represent a promising business. Among the prospective products, it is possible to find tools prepared by law firms, consulting firms, and institutions. For instance, *"Conducting an AI Risk Assessment"* by Bloomberg Law[30] shows the need for a quality management system with written policies, procedures, and instructions, and provides a detailed methodology for AI risk assessments. Similarly, *"Article 43: Conformity Assessment"* by SECURITI.AI[31] focuses on assessing quality management systems and documentation to ensure compliance with the AI Act for high-risk AI systems.

The European Fund and Asset Management Association (EFAMA)'s AI-system Assessment Tool[32] helps firms understand AI regulatory complexities, including compliance with the EU AI Act and the GDPR. This tool also aids in documenting and assessing AI use cases, particularly for high-risk systems.

[25] Edyta Bogucka and others, 'Co-Designing an AI Impact Assessment Report Template with AI Practitioners and AI Compliance Experts' in *AIES '24: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol 7 (AAAI Press 2025).

[26] Dotan Ravit and others, 'Evolving AI risk management: A maturity model based on the NIST AI risk management framework' [2024] arXiv:2401.15229.

[27] Delaram Golpayegani and others, 'AIRO: an ontology for representing AI risks based on the proposed EU AI Act and ISO risk management standards' (IOS Press 2022) 51-65

[28] Jessica Kelly and others, 'Navigating the EU AI Act: A Methodological Approach to Compliance for Safety-Critical Products' in *2024 IEEE Conference on Artificial Intelligence (CAI)* (IEEE 2024) <https://doi.org/10.1109/CAI59869.2024.00179> accessed 15 September 2025.

[29] Simon Chard, Brent Johnson and Daniel Lewis, 'Auditing Large Language Models for Privacy Compliance with Specially Crafted Prompts' [2024] OSF Preprint <https://osf.io/preprints/osf/8tgkx_v1> accessed 11 September 2025.

[30] Arsen Kourinian and Mayer Brown, 'Conducting an AI Risk Assessment' (*Bloomberg Law*, January 2024) <www.bloomberglaw.com/external/document/X3D03D2K000000/data-collection-management-overview-conducting-an-ai-risk-assess> accessed 28 February 2025.

[31] SECURITI.AI, 'Article 43: Conformity Assessment' (2024) <https://securiti.ai/eu-ai-act/article-43/> accessed 28 February 2025.

[32] EFAMA, 'EFAMA's AI-system Assessment Tool' (5 February 2025) <https://www.efama.org/newsroom/news/efama-s-ai-system-assessment-tool> accessed 28 February 2025.

PwC offers an AI Compliance Tool[33] to guide organisations through the AI Act's requirements. It supports risk assessment, classification, and compliance documentation across technical, business, and compliance teams.

Hogan Lovells' AI Act Applicability Assessment[34] and AI HR Systems Compliance are two questionnaires that are available after registration on the website and designed to help organisations verify the AI Act's applicability to concrete use cases and relevant obligations, offering a preliminary view of the impact of the AI Act.

The European AI Scanner[35] is a RegTech tool designed to ensure compliance with the EU AI Act by standardising AI project workflows. It covers data ingestion, model development, deployment, and lifecycle management, providing a comprehensive solution for firms to comply with AI regulations, especially the AI Act.

3 The need for dataset assessment tools

Data is used in AI training through a process known as machine learning, where algorithms learn patterns, correlations, and decision-making rules from large volumes of structured or unstructured data. During training, an AI model based on machine learning is exposed to a dataset comprising examples relevant to the task at hand so that it can identify statistical relationships within that data. For instance, in supervised learning, a model is trained on labeled datasets, where each input (e.g., a medical image) is paired with a correct output (e.g., a diagnosis). The algorithm adjusts its internal parameters iteratively to minimise the difference between its predictions and the correct outputs, a process guided by optimisation techniques, such as gradient descent.

In unsupervised learning, the AI is given unlabeled data and learns to detect hidden patterns or groupings, such as when patients with similar symptoms are clustered. Reinforcement learning, another approach, involves learning optimal behaviours through trial and error, based on feedback from its environment.

Regardless of which method is used, the quality, quantity, and diversity of the training data are critical: biased, incomplete, or imbalanced data can lead the model to learn inaccurate or discriminatory patterns[36]. Thus, data in AI training is not only a technical input, but also a determinant of ethical and operational outcomes, which needs to be curated and evaluated throughout the AI lifecycle.

Datasets are the foundational building blocks of artificial intelligence, shaping the behavior, accuracy, and decision-making capacity of AI systems. Every algorithm learns

[33] PwC, 'AI Compliance Tool' (PwC, 2025) <www.pwc.com/cz/en/sluzby/technologie-a-data/ai-act/ai-compliance-tool.html> accessed 28 February 2025.

[34] Hogan Lovells, 'AI Act Applicability Assessment' (Digital Client Solutions) <https://digital-client-solutions.hoganlovells.com/ai-act-applicability-assessment?utm> accessed 28 February 2025.

[35] European Commission, 'European AI Scanner - RegTech Tool for EU AI Act Compliance' (1 August 2023) <https://futurium.ec.europa.eu/en/european-ai-alliance/best-practices/european-ai-scanner-regtech-tool-eu-ai-act-compliance> accessed 28 February 2025.

[36] Rispoli and others (n 14).

patterns, associations, and predictive models from the data it is exposed to; thus, the quality, scope, and structure of datasets directly influence how an AI system understands and interacts with the world. Data serves not merely as input, but also as a lens through which AI systems interpret social, economic, and medical realities. For example, in healthcare, datasets containing patient histories, clinical notes, and diagnostic images allow AI to detect diseases or recommend treatments. However, if these datasets are incomplete, biased, or unrepresentative of diverse populations, the AI system may generate misleading or harmful outputs. As such, datasets are not neutral artifacts, for they reflect the social and institutional contexts in which they are produced[37]. Recognising this, it is vital to approach data as a site of ethical, technical, and political concern, requiring deliberate curation, validation, and evaluation to ensure that AI systems are trained equitably and reliably across different populations and environments.

The use of data for AI training is classified as a secondary use (commonly referred to as "reuse"): the data is primarily collected for the so-called primary use, and only later is it employed for AI (see Fig 1 below for an overview of the data lifecycle and related biases).

[37] See, among others: Abeba Paullada and others, 'Data and its (dis)contents: A survey of dataset development' (2021) 2(11) Patterns 100357; Emily Denton, Brent Hecht and Lauren Wilcox, 'On the genealogy of machine learning datasets: A critical history of ImageNet' (2021) 8(2) Big Data & Society <https://journals.sagepub.com/doi/full/10.1177/20539517211035955> accessed 15 September 2025; Alexandra Sasha Luccioni and others, 'Position: Measure Dataset Diversity, Don't Just Claim It' [2024] arXiv preprint arXiv:2407.08188; Milagros Miceli, Julian Posada and Tianling Yang, 'Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power?' (2022) 6 Proceedings of the ACM on Human-Computer Interaction 1; Rodrigo R Gameiro and others, 'The Data Artifacts Glossary: a community-based repository for bias on health datasets' (2025) 32(14) Journal of Biomedical Science 1; Shamik Bose Santy and others, 'NLPositionality: Characterizing Design Biases of Datasets and Models' [2023] arXiv preprint arXiv:2306.01943.

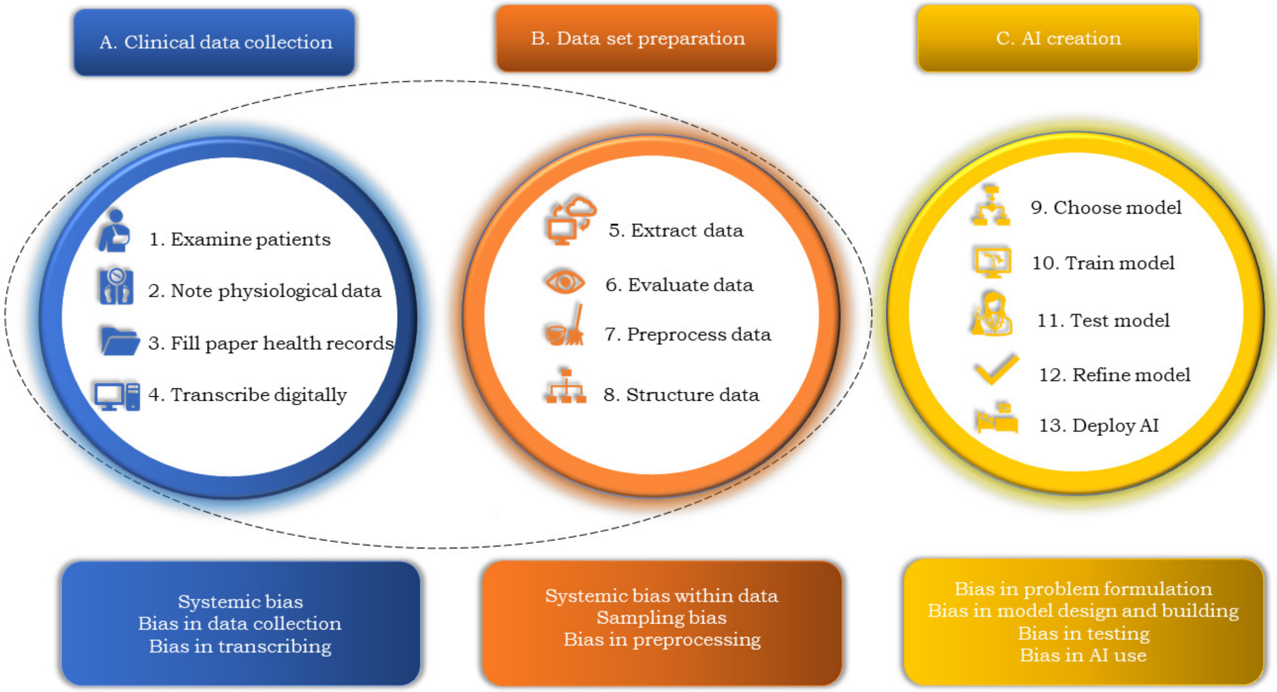

Figure 1 - AI lifecycle and related biases

For instance, in clinical practice, data collection is often performed by subjects with no professional AI knowledge (nurses, MD students, doctors) primarily for a medical purpose (i.e., to deliver healthcare services or to perform clinical trials, though not to create an AI dataset),[38] in a non-standardised way (each hospital has different processes, practices, and IT systems),[39] and not optimised for AI training (data is mostly unstructured, not machine-readable, and often even on paper). The information collected (“raw data”) is relevant for primary use but may not be of sufficient quality for proper secondary use.

When data is transcribed in the electronic patient record (EPR) systems and extracted for AI use, there might be additional quality loss and errors[40]. When medical experts are involved in data labelling, other types of bias occur. On the other hand, AI trainers are not usually experts in healthcare, and they often feed the model with the only data available, pre-processing it solely to boost its performance, without considering if the

[38] Giovanni Tripepi and others, ‘Selection Bias and Information Bias in Clinical Research’ (2010) 115(2) Nephron Clinical Practice 94–99 <https://doi.org/10.1159/000312871> accessed 15 September 2025.

[39] Derek Kyte and others, ‘Systematic Evaluation of the Patient-Reported Outcome (PRO) Content of Clinical Trial Protocols’ (2014) 9(10) PLoS ONE e110229 <https://doi.org/10.1371/journal.pone.0110229> accessed 15 September 2025.

[40] Abimbola A Ayorinde and others, ‘Publication and Related Biases in Health Services Research: A Systematic Review of Empirical Evidence’ (2020) 20(1) BMC Medical Research Methodology 137 <https://doi.org/10.1186/s12874-020-01010-1> accessed 15 September 2025; Augustus A White and others, ‘Self-Awareness and Cultural Identity as an Effort to Reduce Bias in Medicine’ (2018) 5 Journal of Racial and Ethnic Health Disparities 34.

dataset is appropriate, if the quality is good enough, if expert knowledge is missing, or if biases are present.[41]

Additionally, patients and other stakeholders are not involved in the dataset creation phase, and their opinions are not heard despite the efforts to develop inclusive technology.[42] Because of this, it is possible that some biases are overlooked: for example, if women are systemically misdiagnosed, women's data in the datasets will be wrong.[43]

Many errors, biases, unevenness in data, and missing information thus occur when such data is employed. However, to date the main actors involved in AI creation have no guidance on bias mitigation and the concrete application of the GDPR and other Union laws adopted within the Digital Single Market Strategy, being completely unprepared to face the European Data Health Space (EDHS)[44] and the AI Act, among others. As detailed in the literature and in our previous research,[45] clinical datasets might, in the end, be unfair and have poor quality, for they may reflect societal or individual biases, due to having too few examples to represent the target population, missing values in rows or columns, or transcription errors, data loss, and so forth.

Several studies have addressed the pervasive problem of biases in artificial intelligence[46], highlighting how both explicit and implicit biases can significantly impact outcomes in domains such as healthcare.[47] A seminal 2019 study revealed that racial bias embedded in clinical algorithms led to diminished care for patients of colour, demonstrating how such systems can perpetuate systemic disparities.[48] These biases often originate from AI systems trained on data that reflects historically constructed gender and racial hierarchies, thus amplifying inequalities when applied to real-world clinical

[41] SN Stuckless, PS Parfrey and MO Woods, 'The Impact of Misclassification Bias on the Estimation of Relative Effect in Stratified Case-Control Studies: A Simulation Study' (2014) 14 BMC Medical Research Methodology 105.

[42] Rebecca Featherston and others, 'Decision Making Biases in the Allied Health Professions: A Systematic Scoping Review' (2020) 15(10) PLoS One e0240716 <https://doi.org/10.1371/journal.pone.0240716> accessed 15 September 2025.

[43] Lorena Alcalde-Rubio and others, 'Gender Disparities in Clinical Practice: Are There Any Solutions? Scoping Review of Interventions to Overcome or Reduce Gender Bias in Clinical Practice' (2020) 19(1) International Journal for Equity in Health 166 <https://doi.org/10.1186/s12939-020-01283-4> accessed 15 September 2025.

[44] Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847.

[45] C Sessa and others, 'Identifying Bias in Data Collection: A Case Study on Drugs Distribution' (2024) International Joint Conference on Neural Networks (IJCNN) 1; C Jones and others, 'A Causal Perspective on Dataset Bias in Machine Learning for Medical Imaging' (2024) 6(2) Nature Machine Intelligence 138; J Vaughn and others, 'Dataset Bias in Diagnostic AI Systems: Guidelines for Dataset Collection and Usage' (2020) Proceedings of the ACM Conference on Health, Inference and Learning 2; R Daneshjou and others, 'Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping Review' (2021) 157(11) JAMA Dermatology 1362.

[46] Reva Schwartz and others, 'Towards a Standard for Identifying and Managing Bias in Artificial Intelligence' [2022] NIST Special Publication 1270 <https://doi.org/10.6028/NIST.SP.1270> accessed 25 September 2025; Eduard Fosch-Villaronga and Adam Poulsen, 'Diversity and Inclusion in Artificial Intelligence' in Bart Custers and Eduard Fosch-Villaronga (eds), *Law and Artificial Intelligence: Regulating AI and Applying AI in Legal Practice* (Springer 2022).

[47] Leo Anthony Celi and others, 'Sources of Bias in Artificial Intelligence that Perpetuate Healthcare Disparities—A Global Review' (2022) 1(3) PLOS Digital Health e0000022 <https://doi.org/10.1371/journal.pdig.0000022> accessed 15 September 2025.

[48] Ziad Obermeyer and others, 'Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations' (2019) 366(6464) Science 447.

settings. The global deployment of AI technologies, frequently developed in the United States or Europe, and then used in Africa, Asia (with the exception of China, which is a major AI producer and exporter), and Latin America, urgently calls for more inclusive data practices.

There is tension in the field of fair ML between two opposing views on AI fairness. On one side, it is argued that "debiasing"[49] and correcting data and models can address rights violations by ensuring AI systems reflect real-world conditions fairly.[50] The EU's policy, for instance, adheres to this view in the AI Act by mandating that AI datasets must be "relevant, sufficiently representative, and to the best extent possible, free of errors and complete in view of the intended purpose" (AI Act, art 10) to minimise bias.

However, critics counter that the problem lies not in biased models but in using certain AI systems to solve some specific problems in the first place, which – according to them – is an inherently inequitable practice.[51] They argue that debiasing data may reinforce these inequities by reflecting a world that is already structurally biased against marginalised communities. The real issue, as these critics argue, is that some AI systems amplify and perpetuate existing forms of structural discrimination rather than eliminate them. As such, focusing on debiasing AI models or datasets is seen as insufficient or even counterproductive, as it distracts from addressing the root cause of the issue: structural

[49] According to the EDRi Report by Balayn and Gürses (n 5), «Debiasing refers to the application of select methods to address bias by achieving certain forms of statistical parity (e.g. making sure that the accuracy of a recidivism prediction system is similar for Black and White people by rebalancing a training dataset and re-training a machine learning model)».

[50] Anna Leschanowsky, Birgit Popp and Nils Peters, 'Debiasing Strategies for Conversational AI: Improving Privacy and Security Decision-Making' (2023) 2 Digital Society 34; Nicholas Meade, Elinor Poole-Dayan and Siva Reddy, 'An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-Trained Language Models' [2021] arXiv:2110.08527 <https://doi.org/10.48550/arXiv.2110.08527> accessed 11 September 2025; Otávio Parraga and others, 'Debiasing Methods for Fairer Neural Models in Vision and Language Research: A Survey' [2022] arXiv:2211.05617 <https://doi.org/10.1145/3637549> accessed 15 September 2025; Marcus Tomalin and others, 'The Practical Ethics of Bias Reduction in Machine Translation: Why Domain Adaptation Is Better than Data Debiasing' (2021) 23 Ethics and Information Technology 419; Andrés Domínguez Hernández and Vassilis Galanos, 'A Toolkit of Dilemmas: Beyond Debiasing and Fairness Formulas for Responsible AI/ML' in *2022 IEEE International Symposium on Technology and Society (ISTAS)* vol 1 (IEEE 2022); Puspita Majumdar, Richa Singh and Mayank Vatsa, 'Attention Aware Debiasing for Unbiased Model Prediction' in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE 2021) <https://doi.org/10.1109/ISTAS55053.2022.10227133>; Yasaman Yousefi, 'Data Sharing as a Debiasing Measure for AI Systems in Healthcare: New Legal Basis' in *Proceedings of the 15th International Conference on Theory and Practice of Electronic Governance* (ACM 2022); Marta Marchiori Manerba and Riccardo Guidotti, 'Investigating Debiasing Effects on Classification and Explainability' in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (ACM 2022); Clarice Wang and others, 'Do Humans Prefer Debiased AI Algorithms? A Case Study in Career Recommendation' in *27th International Conference on Intelligent User Interfaces* (ACM 2022); Ricards Marcinkevics, Ece Ozkan and Julia Vogt, 'Debiasing Deep Chest X-Ray Classifiers Using Intra- and Post-Processing Methods' (2022) 182 PMLR 504; Carlo Alberto Barbano, Enzo Tartaglione and Marco Grangetto, 'Bridging the Gap between Debiasing and Privacy for Deep Learning' in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE 2021); Ramon Correra and others, 'A Robust Two-Step Adversarial Debiasing with Partial Learning: Medical Image Case-Studies' in *Medical Imaging 2023: Imaging Informatics for Healthcare, Research, and Applications*, vol 12469 (SPIE 2023); Emma AM Stanley, Matthias Wilms and Nils D Forkert, 'Disproportionate Subgroup Impacts and Other Challenges of Fairness in Artificial Intelligence for Medical Image Analysis' in John SH Baxter and others (eds), *Ethical and Philosophical Issues in Medical Imaging, Multimodal Learning and Fusion Across Scales for Clinical Decision Support, and Topological Data Analysis for Biomedical Imaging* (Springer 2022).

[51] Balayn and Gürses (n 5).

injustice. According to this view, debiasing AI cannot overcome societal inequities when the broader system it operates within is fundamentally unequal.

However, despite the efforts of the ML fairness community, Valdivia and others argue that “literature partially fails to show that datafication reinforces racial profiling beyond the creation of racial categories as features” and that “[a] growing scholarship has discussed how datafication is grounded on algorithmic discrimination. However, these debates only marginally address how racialised classification or race categories are enforced through quantification and neglect its political and historical conceptualisation”.[52] In addition, some of the arguments in the ML fairness community can be applied to any technology, not just AI. For example, inequalities in medical technology access in wealthy insurance-based countries remain present not only with regard to AI systems but also to basic medical instruments, such as X-rays, RMN, TC, and ultrasound scans. By funding both a new RMN machine and an AI-based diagnostic tool in a wealthy neighbourhood as a reward system for good performance and low surgery error rate, the decision-makers are increasing the gap against disadvantaged citizens. Therefore, in the absence of a political decision to reallocate resources, the only effective solution to fully eliminate inequalities between those who can access healthcare and those who cannot would be eliminating technology from everyday life *tout court*, so that nobody has an unfair advantage. This solution is not what this article argues, nor is it within its scope to focus on the solution to this issue.

This article’s arguments fully consider the possible harmful conceptualizations of the AI system’s scope: for example, we would consider it to be acceptable to have a correctly trained AI system that recognises malicious cells, but not a tool that would suggest if the patient needed to be treated with chemotherapy, as this decision might be influenced by institutional racism (even in the absence of an AI system). In addition, as shown by our previous research, we adhere to Cynthia Rudin’s and other authors’ positions,[53] according to which black boxes should not be used in medicine, and we advocate for a “right to technical interpretability”.[54] When white boxes are involved, many problematic issues of AI described in the EDRi report are eliminated or mitigated.

This article takes, however, the above-mentioned stream of research very much into consideration. We argue that, while a correct, statistically sound, and fair dataset should be the prerequisite for any AI application in healthcare, it is a condition necessary but not sufficient – alone – to eliminate biases worldwide, especially in countries rooted in a colonialist past. For this reason, the article does not aim to promote debiasing in AI but

[52] Ana Valdivia and Martina Tazzioli, ‘Datafication Genealogies beyond Algorithmic Fairness: Making up Racialised Subjects’ in *FAccT ’23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2023).

[53] Chiara Gallese, ‘The AI Act Proposal: A New Right to Technical Interpretability?’ in Paulina Kowalicka (ed), *Internet Law and Digital Society: An International Overview* (Milano University Press 2025).

[54] Elena Falletti and Chiara Gallese, ‘Credit Scoring Judicial Review between the Court of Justice of the European Union and Comparative Case Law’ (2024) 3 Media Laws 92.

rather to incorporate critical studies in the epistemology of datafication of patients, advancing the debate on structural injustice in the use of AI.

In recent years, the notion that only the use of data is political has been challenged: the way data is collected and processed represents a political choice.[55] The apparatus of dataset production needs to be studied, as noted in the literature.[56] To do so, we do not rely on a model/system-centric view but take a socio-technical approach to the issue, considering power relations along with technical issues.

Addressing these issues necessitates a structural shift in how datasets are conceived and validated. Drawing on Michel Foucault's theories[57] and Kitchin and Lauriault's notion of "data assemblages",[58] AI datasets must be recognized not merely as technical constructs but as socio-political products embedded in power relations. The datafication process, shaped by unequal access to care and resources, reinforces marginalisation by encoding and legitimising hierarchical distinctions through algorithmic design. Inadequate representation in datasets compromises diagnostic accuracy and patient safety, threatening public trust in AI systems. Thus, ethical AI development demands a critical examination of the data infrastructures that shape clinical intelligence, with greater transparency, accountability, and participatory input from diverse and vulnerable communities.

Given the critical role that datasets play in shaping AI outcomes, the development and implementation of robust data assessment tools is essential. These are necessary to systematically evaluate the provenance, composition, and representativeness of data used to train AI systems. Without standardised mechanisms for auditing datasets, biased or incomplete data can go undetected, embedding structural inequalities into algorithmic outputs.

Data assessment tools help identify gaps in demographic representation, highlight potential sources of skewed labelling or annotation practices, and assess the contextual relevance of data across different populations and regions. Especially in high-stakes fields like healthcare, where flawed data can result in misdiagnoses or inappropriate treatments, such tools serve as a critical checkpoint to safeguard patient welfare.

[55] Chiara Gallese, 'Redefining Anonymization: Legal Challenges and Emerging Threats in the Era of EHDS' in *Enabling and Safeguarding Personalized Medicine* (Springer 2025).

[56] Remi Denton and others, 'Bringing the People Back in: Contesting Benchmark Machine Learning Datasets' [2020] arXiv:2007.07399 <https://doi.org/10.48550/arXiv.2007.07399> accessed 15 September 2025; Eun Seo Jo and Timnit Gebru, 'Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning' in *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (ACM 2020); Michael Katell and others, 'Toward Situated Interventions for Algorithmic Equity: Lessons from the Field' in *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (ACM 2020); Os Keyes, 'The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition' (2018) 2 (CSCW) Proceedings of the ACM on Human-Computer Interaction 88.

[57] Michel Foucault, Michel Senellart (ed), *The Birth of Biopolitics: Lectures at the Collège de France, 1978-1979* (Palgrave Macmillan 2008).

[58] Rob Kitchin and Tracey P Lauriault, 'Towards Critical Data Studies: Charting and Unpacking Data Assemblages and Their Work' in Jim Eckert, Andrew Shears and Josef Thatcher (eds), *Thinking Big Data in Geography: New Regimes, New Research* (University of Nebraska Press 2018).

Moreover, as AI technologies increasingly cross jurisdictional boundaries, transparent and interoperable assessment frameworks ensure that ethical standards are upheld globally, not just in the regions where the AI was developed.

4 The FAnFAIR tool

This tool develops a semi-automated framework to evaluate the fairness of datasets employed in machine learning applications, especially in sensitive domains such as healthcare. The method integrates statistical indicators with normative and legal considerations to generate a composite fairness score, thereby enabling principled decision-making regarding data use. Recognizing the legal implications of biased or non-compliant data, the framework draws from EU data protection law, principles of algorithmic fairness as drawn in the relevant literature, and established ethical guidelines to operationalise fairness assessment.

FAnFAIR evaluates datasets across six core dimensions: balance (distribution of target classes), numerosity (ratio of data points to features), unevenness (presence of statistical outliers), compliance (conformity with legal and ethical standards), quality (domain expert assessment), and incompleteness (extent of missing data). Each dimension is formalised as a linguistic variable within a fuzzy logic (FL) system, which permits reasoning under uncertainty and vagueness, conditions often encountered in legal and data governance contexts. Fuzzy logic is applied using a Takagi–Sugeno inference model,[59] implemented through the Python-based Simpful library.[60] Each variable is evaluated through membership functions and mapped to fairness-related rules structured in a fuzzy rule base.

The compliance dimension uniquely captures multi-jurisdictional legal concerns, including conformity with the GDPR, copyright law, medical and bioethical standards, and non-discrimination law. Because compliance involves jurisdiction-specific legal obligations, it is not automatically derived from the data; instead, it must be manually set by the user based on a detailed legal checklist encompassing lawful basis, consent, anonymization, data minimisation, and stakeholder accountability, among others.

The resulting fairness score, expressed on a 0–100 scale, synthesises the output of all fuzzy rules and reflects both technical data characteristics and legal-ethical considerations. In our first work, the tool was applied to two public datasets from the UCI Machine Learning Repository to illustrate its functionality and generate preliminary insights. These applications demonstrate the viability of FAnFAIR as a transparency-

[59] Tomohiro Takagi and Michio Sugeno, 'Fuzzy Identification of Systems and Its Applications to Modeling and Control' (1985) 15(1) IEEE Transactions on Systems, Man, and Cybernetics 116.

[60] Simone Spolaor and others, 'Simpful: A User-Friendly Python Library for Fuzzy Logic' (2020) 13 International Journal of Computational Intelligence Systems 1687.

enhancing instrument to assess data fairness prior to AI deployment, aligning technological design with both ethical principles and legal requirements.

Fuzzy Logic (FL) represents a form of many-valued logic specifically developed to manage imprecision and uncertainty, characteristics often inherent in complex real-world systems. Unlike classical binary logic, which permits only strict true/false evaluations, FL enables decision-making based on degrees of truth. At its core is the *fuzzy set*, an extension of traditional set theory that allows an element to possess a *membership degree* between 0 and 1. This approach permits partial inclusion in multiple sets simultaneously, thereby capturing nuances that conventional logic systems overlook.

To quantify these membership degrees, *membership functions* (MFs) are employed. A commonly used MF is the triangular function, which offers intuitive modelling of gradual transitions between fuzzy categories. In our study, we adopt normalised triangular fuzzy sets, ensuring that the cumulative membership values across the defined domain equal one. This normalisation supports a coherent assessment of input variables while retaining the capacity for refinement through more sophisticated functions if required.

Fuzzy sets are used to define *linguistic variables*, that is, variables articulated through natural language expressions such as “low”, “moderate”, or “high.” For instance, a variable like “room temperature” may be described by the linguistic terms “cold” or “warm,” each corresponding to its own fuzzy set. A specific temperature value, like 25 ° C, may then belong simultaneously to both sets with varying degrees of membership, for example, minimally to “cold” and significantly to “warm.”

These linguistic variables serve as the foundation for *fuzzy rules*, which operate in the form of conditional logic, typically expressed as:

IF *X* is *a* **THEN** *Y* is *b*,

where *X* and *Y* are linguistic variables, and *a* and *b* are associated linguistic terms. Such rules permit partial satisfaction; each rule is evaluated on a scale from 0 (not satisfied) to 1 (fully satisfied). A typical fuzzy inference system contains multiple such rules to address different conditions and nuances in the data.

In our implementation, we employed a 0-order *Takagi-Sugeno* inference system,[61] which synthesises the outcome of all applicable rules through weighted averaging. The output is computed using the formula:

$$output = \frac{\sum_{i=1}^{N} w_i z_i}{\sum_{i=1}^{N} w_i},$$

where *N* is the number of rules, w_i denotes the degree of satisfaction of the *i*-th rule, and z_i is the output value associated with that rule. The rule-based system was constructed

[61] Takagi and Sugeno (n 60).

using the *Simpful* Python library, a specialised tool designed to streamline the development of fuzzy inference models. Simpful supports natural-language rule encoding and allows for flexible and complex membership functions. For the purposes of this research, we configured the library to use the 0-order Takagi-Sugeno model, aligning with our need for interpretability and precision in fairness assessments of data sets.

4.1 Statistical considerations

We identified the following statistical features that we consider important in a dataset: balance, unevenness, and incompleteness.

A dataset may be considered *balanced* when all relevant classes or categories are proportionately represented across its entries. This concept is of central importance in both technical and legal discussions surrounding algorithmic fairness, as a lack of balance can lead to systemic bias in AI outputs.[62] Specifically, when one class is significantly underrepresented, machine learning models may exhibit skewed behaviour, undermining their generalisability and accuracy. Such an imbalance has been empirically linked to prejudicial outcomes, including disparities in error rates and misclassification - issues that disproportionately affect marginalised groups and may give rise to legal and ethical concerns under anti-discrimination and data protection regimes.

To illustrate, consider a medical dataset used to train an AI model to classify patients as either "healthy" or "non-healthy." If the dataset includes a vast majority of diseased cases but only a few examples of healthy individuals, the resulting model is likely to exhibit diminished accuracy in identifying healthy patients. This not only compromises clinical reliability but may also infringe upon the rights of individuals by perpetuating unequal treatment, especially if protected characteristics correlate with underrepresented classes.

To quantify balance in formal terms, we calculate a *normalised balance score*. First, the frequency of each class is determined and divided by the total number of instances to generate a class ratio vector. The standard deviation of this vector, denoted by σ, reflects the extent of distributional disparity. A perfectly balanced dataset would have a standard deviation of zero. Conversely, in the worst-case scenario, where all data points belong to a single class, the standard deviation reaches a maximum value, denoted σ*. The balance score is therefore defined as:

$\text{Balance}=1-\sigma/\ \sigma*$

This metric ranges from 0 (complete imbalance) to 1 (perfect balance) and provides a foundational basis for assessing the representational fairness of a dataset. In contexts where algorithmic decisions affect fundamental rights, such as healthcare access,

[62] Natalia Criado and Jose M Such, 'Digital Discrimination' in Karen Yeung and Martin Lodge (eds), *Algorithmic Regulation* (OUP 2019); Philippe Burlina and others, 'Addressing Artificial Intelligence Bias in Retinal Diagnostics' (2021) 10(2) Translational Vision Science & Technology 13; Taeuk Jang, Feng Zheng and Xiaoqian Wang, 'Constructing a Fair Classifier with Generated Fair Data' (2021) 35(9) Proceedings of the AAAI Conference on Artificial Intelligence 7908.

employment, or criminal justice, this measure may serve as a preliminary but crucial safeguard against discriminatory data practices.

Numerosity, in the context of dataset evaluation, refers to the quantity of data instances relative to the number of features, formally expressed as the ratio $\alpha = S / D$, where S denotes the number of instances and D the number of features. This ratio is critically important for ensuring that machine learning models trained on a dataset can produce valid and generalizable inferences. A dataset with insufficient numerosity may lead to overfitting, where the model captures noise rather than meaningful patterns, resulting in outputs that are statistically misleading and potentially harmful, especially in contexts involving high-stakes decision-making.

To illustrate, consider a dataset in which the input features bear no real relationship to the target outputs. Ideally, a machine learning model would not perform well on such data, reflecting the lack of substantive correlation. However, if the number of features exceeds the number of instances, it becomes statistically possible, even trivial, for an algorithm to perfectly "fit" the data through random associations. Empirical studies confirm that linear systems of equations, when overparameterised in this way, can be solved with probability one, even if based on entirely random inputs. This mathematical phenomenon allows for the appearance of model accuracy where none exists, undermining trust and violating principles of scientific validity and procedural fairness. In legal and ethical terms, deploying such a model could be interpreted as negligent, or even discriminatory, if it systematically fails to generalise across different populations or contexts, especially if those populations are protected by anti-discrimination laws. To guard against such risks, best practices in statistical modelling recommend that datasets maintain a minimum threshold of numerosity, commonly approximated as $S \geq 10 \times D$.[63] This guideline provides a safeguard against learning from coincidental or spurious patterns that are not representative of the real-world phenomena the model seeks to capture.

Although more advanced assessments of dataset sufficiency can be performed using the Vapnik–Chervonenkis (VC) dimension, which relates to the complexity of the hypothesis space, this requires prior knowledge of the model architecture, an assumption that is often impractical at the data collection stage. Consequently, the simplified numerosity ratio α serves as a pragmatic and legally prudent proxy for evaluating whether a dataset is sufficiently robust to support ethical and lawful AI development.

For computational purposes, this feature is defined over the universe of discourse [0, 10], and the numerosity score is calculated as:

$$\text{Numerosity} = \min(10, \alpha)$$

where a value of $\alpha \geq 10$ reflects an optimally sized dataset suitable for model training. This threshold contributes to the overall fairness assessment of a dataset and supports compliance with emerging regulatory standards for trustworthy and non-discriminatory AI.

[63] FE Harrell, *Regression Modeling Strategies* (2nd edn, Springer 2015).

Incompleteness refers to the presence of missing values, also referred to as "holes", within a dataset, particularly with respect to relevant attributes or variables. This phenomenon is common in real-world data collection, especially in high-stakes domains such as healthcare, where individual patients may lack information for specific physiological parameters.

The presence of such gaps presents both statistical and legal challenges. From a modelling standpoint, missing data can reduce accuracy, impair generalisability, and lead to erroneous outputs. From a legal and ethical perspective, it may compromise the integrity of informed decision-making, skew representation across demographic groups, and violate principles of fairness and due diligence under data governance laws.

Depending on the extent and distribution of missing values, researchers may choose to apply imputation techniques (e.g., statistical estimation of missing values) or to exclude affected cases entirely. The choice between these approaches must be made carefully, balancing statistical robustness with respect for individual dignity and data minimisation principles enshrined in data protection regimes such as the GDPR.

In practice, three common types of incompleteness can be observed:

1. Variable-level incompleteness – where certain columns (features) contain missing values. This typically occurs when certain tests or measurements are not universally administered across all individuals.
2. Instance-level incompleteness – where specific rows (individuals) lack multiple values, often due to medical fragility or procedural limitations.
3. Randomly distributed incompleteness – where missing values appear without systematic pattern across the entire dataset.

For purposes of computational simplicity, this study assumes the third form of incompleteness. The degree of incompleteness is assessed by calculating the ratio of missing to total values across the dataset. We define a binary matrix I ∈ {0,1}^{S × D}, where S is the number of samples and D the number of features. Each element i_{d,s} is defined as:

i_{d,s} =
1 if the value is missing (NaN),
0 otherwise

The incompleteness score is then calculated as the proportion of missing entries:

$$incompleteness = \frac{\sum_{d=1}^{D}\sum_{s=1}^{S} i_{s,d}}{S \times D}.$$

This metric yields a normalised score ranging from 0 (indicating a complete dataset) to 1 (indicating that all values are missing). As with other fairness-related dimensions, high levels of incompleteness may signal the need for closer scrutiny of data collection protocols, legal compliance, and potential adverse impacts on underrepresented or vulnerable populations.

4.2 Qualitative Considerations

We identified two qualitative features that we consider essential for a dataset to be fair: quality and compliance.

Quality refers to the substantive reliability and interpretability of the data, as assessed by a qualified domain expert. This variable is particularly salient in fields such as medicine, where input data may consist of images, physiological signals, or clinical annotations that require expert interpretation to determine their usability and clarity. For instance, the recognisability of a radiographic image or the audibility of a recorded voice sample is critical in determining whether the data can support accurate and ethically sound AI inferences.

The quality of data is inherently context-sensitive and cannot be automatically measured using computational methods alone. Instead, domain experts must evaluate the dataset in accordance with the latest standards of their field, factoring in both scientific rigour and practical relevance. This expert-driven evaluation is crucial because data collection processes are frequently decentralised or inconsistent, involving individuals with varying levels of expertise, such as medical trainees, technicians, or crowdsourced contributors. As a result, data may be incomplete, mislabelled, or poorly annotated, undermining the reliability of any models trained on it.

Data quality also includes the design and execution of the annotation process. Large-scale data labelling, especially through crowdsourcing, poses significant challenges to label accuracy, especially in high-stakes environments like healthcare. In such contexts, systematic quality assurance mechanisms, such as assessing annotator performance and measuring inter-rater reliability, become essential. For example, collaborative initiatives between clinicians and AI researchers have demonstrated the importance of iterative quality control in medical image annotation.[64]

In our study, we adopt a holistic and manual approach to evaluating data quality. The *FAnFAIR* framework includes a set quality method, which allows domain experts to assign a quality score based on a broad array of considerations. These may include the clarity of the annotation scheme, the comprehensiveness of labelling instructions, the interpretability of features, and alignment with professional standards. This approach is conceptually aligned with best practices from fact-checking and media accountability organisations, which similarly rely on expert-driven, context-aware quality metrics.[65]

As with legal compliance, the quality variable is not computable through automated means. Indeed, it requires reflective judgment from those with domain-specific expertise, ensuring that AI systems are trained on datasets that are not only statistically robust but also ethically and professionally sound.

[64] Beverly Freeman and others, 'Iterative Quality Control Strategies for Expert Medical Image Labeling' (2021) 9 Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 60.

[65] 'Website Rating Process and Criteria' (*NewsGuard*) <www.newsguardtech.com/ratings/rating-process-criteria/> accessed 11 September 2025.

Compliance is a feature composed of multiple sub-features. It refers to the conformity of a dataset and its underlying data practices with applicable legal and regulatory frameworks. These include, but are not limited to, data protection laws, intellectual property rights, medical and clinical regulations, and anti-discrimination laws. In this study, our compliance model draws primarily on European Union law and the civil law tradition, although it is adaptable to other jurisdictions and normative systems. We acknowledge that the notion of fairness is culturally contingent and that different societies may place emphasis on varying legal, moral, and social standards.

While legal compliance alone does not exhaust the ethical evaluation of data, it constitutes a foundational threshold. The *lawfulness* criterion is one of the three pillars of *Trustworthy AI* articulated by the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG).[66] The use of unlawfully collected or processed data in AI systems violates both legal obligations and ethical norms, rendering such systems fundamentally flawed from the standpoint of legitimacy and fairness[67].

In our framework, we operationalise compliance through five key dimensions, each reflecting essential legal and ethical considerations:

1. **Data Protection Law**: This includes compliance with principles such as anonymization, establishing a lawful basis for data processing, conducting Data Protection Impact Assessments (DPIA), Legitimate Interest Assessments (LIA), Transfer Impact Assessments (TIA), and Re-use Impact Assessments (RIA), as well as adhering to contractual obligations, data minimization principles, data retention limitations, the rights of data subjects, and transparency duties.
2. **Copyright Law**: Ensuring that data collection and utilisation respect applicable intellectual property regimes, licensing terms, and authorial rights.
3. **Medical Law**: Including requirements for informed consent, ethical review by institutional bodies, and adherence to clinical and diagnostic regulations specific to health-related data.
4. **Non-Discrimination Law**: Affirming the principles of equality and non-discrimination and the prohibition of discrimination on grounds such as gender, ethnicity, disability, or socioeconomic status, in accordance with both national and international legal instruments.
5. **Ethics**: Encompassing broader normative standards such as accountability, respect for human dignity and autonomy, traceability of data practices, stakeholder involvement, risk assessment, and evaluation of societal impact.

These dimensions serve as evaluative checkpoints for assessing whether a dataset is fit for ethical and lawful deployment in AI systems. Importantly, while some elements can

[66] High-Level Expert Group on Artificial Intelligence (AI HLEG), 'Ethics Guidelines for Trustworthy AI' (European Commission 2019) 6.

[67] B Buruk, PE Ekmekci and B Arda, 'A Critical Perspective on Guidelines for Responsible and Trustworthy Artificial Intelligence' (2020) 23(3) Medicine, Health Care and Philosophy 387.

be technically validated (e.g., the presence of consent documentation), others require legal or expert judgment tailored to the jurisdiction and context in which the dataset is used.

The first and arguably most foundational criterion for evaluating dataset compliance concerns adherence to data protection legislation, especially within the European framework. Instruments such as the GDPR, the Council of Europe's Convention 108+, and relevant national regulations establish binding obligations for the lawful processing of personal data. Importantly, these legal standards may also apply to anonymised datasets, particularly where re-identification risks remain non-negligible or where anonymisation is incomplete.

In alignment with GDPR principles and associated jurisprudence, a dataset intended for AI development must satisfy the following key elements to be deemed compliant:

- **Anonymisation**: Where data anonymisation is undertaken, it must be sufficiently robust to prevent re-identification by any reasonably foreseeable means, including linkage with external datasets.
- **Legal Basis**: The collection and use of personal data must be grounded on a lawful basis as defined under Article 6 GDPR, such as informed consent, contractual necessity, legal obligation, vital interests, public interest, or legitimate interest.
- **Data Protection Impact Assessment (DPIA)**: For data processing operations likely to result in a high risk to individual rights and freedoms, a DPIA must be conducted in accordance with Article 35 GDPR, as further detailed by guidance from the European Data Protection Board and national data protection authorities.
- **Legitimate Interest Assessment (LIA)**: If the processing relies on legitimate interests, a thorough balancing test must confirm that such interests are not overridden by the data subjects' fundamental rights and that the processing is necessary and proportionate to pursue those interests (Article 6(1)(f) GDPR).
- **Transfer Impact Assessment (TIA)**: Where data is transferred outside the European Economic Area, whether for storage or processing, an evaluation of the legal and security implications of the transfer must be carried out to ensure adequacy and compliance with Chapter V of the GDPR.
- **Re-use Impact Assessment (RIA)**: Pursuant to the purpose limitation principle, data can only be reused for purposes compatible with the original collection context. Compatibility assessments are necessary to confirm lawful secondary use.
- **Contracts**: In cases involving joint controllership or outsourcing (e.g., through processors), data sharing must be formalised through appropriate legal agreements that clearly define roles and responsibilities, as required by Article 28 GDPR.
- **Data Minimisation**: Only data strictly necessary for the specified processing purposes may be collected or retained (Article 5(1)(c) GDPR);
- **Data Retention**: Personal data must be retained only for the duration necessary to fulfil its intended purpose and must be securely deleted thereafter (Article 5(1)(e) GDPR).

- **Data Subjects' Rights**: Data subjects must be able to exercise their rights effectively, including rights of access, rectification, erasure, restriction, objection, and data portability, as set out in Articles 12–22 GDPR.
- **Transparency Obligations**: The principles of transparency and accountability require clear, accessible, and comprehensive disclosures about data processing practices, including through privacy notices and consent mechanisms.

Compliance with these requirements is not only a legal mandate but a foundational condition for ethical AI. Failure to address them may result in legal liability, reputational harm, and violations of individual rights, especially in AI applications involving vulnerable populations or sensitive data categories.

The second compliance criterion pertains to intellectual property law, particularly copyright and licensing. The legal status of a dataset, including its collection, use, and distribution, is governed by national and international IP laws. The specific legal provisions applicable to a dataset depend on several factors, including the jurisdiction in which the data was created, the intended purposes of use (e.g., research, commercial deployment), and any contractual agreements governing rights and limitations between the data owner (or creator) and the data user. For the data to be lawfully processed and repurposed, both its initial collection and its downstream use must conform to applicable licensing terms and copyright restrictions. Failure to do so may constitute a breach of statutory rights or contractual obligations, exposing parties to legal liability and rendering the AI outputs ethically and legally tainted.

The third criterion addresses compliance with medical and health law, which varies substantially across jurisdictions. Key requirements include obtaining valid informed consent, respecting age-of-consent thresholds, complying with national ethical review standards, and observing rules concerning the reuse of medical data. Some legal systems impose specific conditions on the handling of data relating to deceased individuals, such as restrictions on post-mortem research or continued privacy protections. The heterogeneity of national frameworks means that any cross-border or transnational use of medical data must be accompanied by a careful legal compatibility assessment, especially when such data is fed into AI systems used in clinical decision-making.

The fourth criterion pertains to non-discrimination law and the broader legal obligation to safeguard the fundamental rights and freedoms of data subjects, particularly those belonging to marginalised or vulnerable populations. Systemic inequities in data collection practices, such as the underrepresentation or misclassification of women, ethnic minorities, persons with disabilities, or neurodivergent individuals, can lead to datasets that reinforce structural bias and perpetuate disparate outcomes. For example, inconsistent documentation of medical symptoms across racial or gender lines can result in biased AI models that offer inferior diagnostic performance for these groups. The fairness of AI systems, especially in healthcare, must therefore be evaluated not only at the modelling stage, but also from the very inception of the data lifecycle. Legal and ethical safeguards must be embedded from the point of collection to ensure that marginalised communities are not disproportionately harmed or excluded.

The fifth and final compliance criterion concerns a set of overarching ethical principles that should guide data practices from the earliest stages of collection. While certain ethical duties are codified in law, such as Article 5(2) GDPR's accountability obligation, many go beyond statutory requirements, offering normative guidance for responsible data governance, even in contexts where personal data or specific legal protections may not apply. These principles are essential to fostering trust, legitimacy, and fairness in the deployment of AI systems, and include:

- **Accountability**: Rooted in both legal frameworks and ethical theory, accountability entails that a clearly designated individual or entity is responsible for decisions, actions, and outcomes arising from data collection and usage. This responsibility applies irrespective of whether personal data is involved.
- **Dignity and Self-Determination**: Respect for human dignity requires that individuals have meaningful control over how their data is used, even where legal consent has been obtained. For example, in the case of deceased persons, their previously expressed wishes should be honoured to the extent possible.
- **Traceability**: Ethical assessments must be documented in a transparent and auditable manner. It should be possible to ascertain when, how, and by whom key ethical decisions were made throughout the data lifecycle.
- **Stakeholder Involvement**: Affected individuals and communities, especially those who bear the highest risks of AI deployment, should be engaged in the development process. Their perspectives must be solicited, considered, and reflected in the design and governance of data systems.
- **Risk Assessment**: Ethical review should include a forward-looking evaluation of the potential harms and injustices that may result from the use of the dataset. These assessments should be conducted proactively, rather than reactively.
- **Societal Impact**: The broader societal consequences of data usage, including its effects on public trust, inequality, and systemic discrimination, must be carefully assessed and addressed through appropriate safeguards and mitigation strategies.

Given the jurisdictional variability and contextual specificity of these considerations, the compliance variable cannot be computed automatically. Instead, users of the FAnFAIR framework are required to manually assign compliance values via the set compliance method. This method takes a dictionary object in which each of the five compliance categories—data protection law, intellectual property law, medical law, non-discrimination law, and ethics—is assigned a Boolean value indicating whether the dataset satisfies the relevant criterion. This approach preserves legal fidelity while allowing for contextual flexibility in the assessment of fairness and legality.

In alignment with the regulatory landscape of artificial intelligence in the European Union, Fundamental Rights Impact Assessments (FRIAs) have emerged as a central requirement under Article 27 of the AI Act.[68] For AI systems classified as *high-risk*, including those used in healthcare, education, law enforcement, and employment, the AI Act mandates that providers conduct an assessment of the system's impact on fundamental rights prior to placing the system on the market or putting it into service. This obligation is grounded in the EU Charter of Fundamental Rights[69] and aims to ensure that AI does not infringe upon rights such as dignity, non-discrimination, privacy, freedom of expression, and access to justice.

A FRIA requires not only the identification of potential harms to individual rights but also a proactive evaluation of systemic and structural risks, including how algorithmic processes might perpetuate inequality, obscure accountability, or undermine democratic oversight. It must include meaningful stakeholder consultation, clear documentation of risk mitigation strategies, and integration with broader governance and accountability mechanisms. Importantly, the FRIA must be maintained as a living document and updated throughout the AI system's lifecycle. In the context of dataset evaluation, the FRIA is aligned with lawful data collection, ethical curation, and fairness-aware design, particularly when datasets include information about vulnerable or historically marginalised populations. The inclusion of FRIA procedures within the FAnFAIR framework thus represents a necessary evolution toward aligning technical practices with constitutional values and legal obligations under EU law.

Table 1 shows how our tool helps comply with the AI Act.

FAnFAIR Dimension	AI Act Article	How FAnFAIR ensures compliance
Balance	Art. 10(3)(a)	Ensures statistical representativeness in training, validation, and testing datasets to prevent or mitigate bias.
Compliance	Art. 10(4) and (5)	Embeds checks for Fundamental Rights Impact Assessment and ensures adherence to legal frameworks like GDPR, non-discrimination law, and other relevant legal instruments.
Numerosity	Art. 10(1)(c) and (e)	Ensures that training, validation and testing data sets is subject to data governance and management practices appropriate for the intended purpose of the high-risk AI system, including data enrichment and an assessment of the availability, quantity and suitability of the data sets.
Unevenness	Art. 10(1)(e)	Ensures that training, validation and testing data sets is subject to data governance and management practices appropriate for the

[68] Alessandro Mantelero, 'The Fundamental Rights Impact Assessment (FRIA) in the AI Act: Roots, Legal Obligations and Key Elements for a Model Template' (2024) 54 Computer Law & Security Review 106020.
[69] Andrea Cosentini and others, 'Assessing the Impact of Artificial Intelligence Systems on Fundamental Rights' [2025] SSRN https://dx.doi.org/10.2139/ssrn.5168579 accessed 15 september 2025.

		intended purpose of the high-risk AI system, including an assessment of the availability, quantity and suitability of the data sets.
Quality	Art. 10(1)	Ensures that training, validation and testing data sets meet the quality criteria of paragraphs 2 to 5.

Table 2 shows how our tool compares with model cards and data sheets.

Function	**Model Cards**	**Data Sheets**	**FAnFAIR**
Model description	Must include this information	Data Sheets only consider data, not the model	Our tool is agnostic towards the model
Intended use	Must include this information	Data Sheets only consider data, not the model	This is part of the compliance assessment
Features used in the model	Must include this information	Data Sheets only consider data, not the model	Our tool only addresses the dataset, not the model
Metrics	Must include this information	Data Sheets only consider data, not the model	Our tool only addresses the dataset, not the model
Data used for training & evaluation	Must include this information	Must include this information	Must include this information
Limitations	Must include this information	Not included	Not included
Ethical Considerations	Must include this information	Not included	Must include this information
Dataset overview & example of data points	Not included	Must include this information	Must include this information
Data source collection, validation, transformations and sampling methods	Not included	Must include this information	Included in the compliance feature
Sensitive attributes	Not included	Must include this information	Included in the new version
Training and evaluation methods	Not included	Must include this information	Our tool only addresses the dataset, not the model
Intended and extended use of data	Not included	Must include this information	Must include this information

5 Conclusions

The increasing legal, ethical, and technical scrutiny placed on artificial intelligence by the European Union's AI Act calls for the creation of robust dataset assessment tools. Assessing datasets for fairness is essential to promoting equity and social justice, above all for marginalised groups who are often underrepresented, misrepresented, or systematically excluded from data collection and technological design.

In many AI systems, biased datasets serve as the foundation upon which models are trained, thereby encoding and amplifying existing social inequalities. If these datasets do not reflect the lived experiences, linguistic variations, health conditions, or socioeconomic realities of marginalized populations, such as racial and ethnic minorities, women, people with disabilities, LGBTQ+ individuals, or economically disadvantaged communities, then the resulting AI outputs risk perpetuating harm. These harms may include misdiagnosis in healthcare, exclusion from financial services, biased hiring practices, or surveillance and policing disparities.

Dataset fairness assessment ensures that such groups are not only included but also accurately represented in AI systems, mitigating the risk of discrimination and enabling equitable access to benefits like improved healthcare, education, and public services. Beyond individual impact, fair datasets strengthen public trust in AI by aligning technological development with principles of democracy, transparency, and human rights. They also promote innovation by ensuring that AI models are robust, generalizable, and ethically sustainable across diverse populations. Ultimately, fairness in datasets is an acknowledgement that data, as a reflection of human systems, must be assessed and evaluated to ensure it does not reinforce injustice but instead advances inclusion and equality.

As demonstrated throughout this article, the FAnFAIR framework provides a concrete, semi-automated solution for evaluating the fairness and regulatory compliance of datasets used in AI development. By combining fuzzy logic with domain expertise and legal criteria, FAnFAIR operationalises abstract regulatory requirements into a transparent, reproducible methodology for dataset governance. Importantly, the framework is not limited to statistical metrics but incorporates qualitative judgments regarding data quality and compliance with legal obligations under data protection, intellectual property, medical, and anti-discrimination laws. It also integrates ethical principles such as dignity, accountability, and stakeholder inclusion, and aligns with the AI Act’s requirement for Fundamental Rights Impact Assessments.

Rather than presenting debiasing as a silver bullet, this paper advocates for a broader epistemological and structural critique of data practices, situated within postcolonial and feminist theoretical frameworks. Our contribution thus expands the conversation from compliance to justice, highlighting how datasets are shaped by power relations and institutional asymmetries. By offering both theoretical reflection and a practical tool, this work aims to support developers, regulators, and scholars in designing AI systems that are not only lawful, but also fair, inclusive, and socially responsible.

In addition, we want to highlight that discarding a dataset when it is found to be biased is a critical safeguard in the development of ethical and legally compliant AI systems. A biased dataset that reflects historical discrimination, underrepresents certain populations, or encodes structural inequalities, can lead to harmful and unjust outcomes when used to train machine learning models. These harms may include misdiagnoses in healthcare, unfair allocation of resources, discriminatory treatment in employment or credit scoring, or the reinforcement of societal stereotypes. In high-risk AI systems, such as those governed by the EU AI Act, the use of a flawed dataset may not only undermine the fairness and accuracy of the system, but also breach fundamental rights and legal obligations, including those related to data protection, non-discrimination, and human rights.

Continuing to use a dataset despite evidence of significant bias, especially when that bias cannot be adequately mitigated through rebalancing, imputation, or filtering, risks legitimising and perpetuating structural injustices. Additionally, it may expose developers, deployers, and regulators to legal liability and reputational damage. In such cases, discarding the dataset is not a failure of innovation but a responsible and necessary decision aligned with ethical design principles and regulatory compliance. It reflects a commitment to harm prevention, respect for affected individuals and communities, and integrity in the deployment of AI technologies. Choosing not to proceed with a dataset that fails fairness thresholds ultimately protects both societal interests and AI systems developers.