AI Fairness-by-Design Multi-Stakeholder Methodology

-

A Comprehensive Framework for Fair AI Design and Development

AEQUITAS
unbias AI

ANNEX IV: Fair DATA Collection, Governance and Management (FDCGM)

Fair-by-Design Sub-methodology

Abbreviation	Meaning
AFF	Affectees
AIU	AI Users
DDM	Development Decisionmakers
DE	Domain Experts
EGTAI	Ethics Guidelines for Trustworthy AI
FbD	Fair-By-Design
FDCGM	Fair Data Collection, Governance, and Management
FMM	Fair Model Methodology
FOIM	Fair Output Interpretation Methodology
FRIA-F	Fundamental Rights Impact Assessment for Fairness
GDM	Governance Decisionmakers
SIM	Stakeholder Identification Methodology
TAIRA	Trustworthy AI Readiness Assessment
MAP	Multistakeholder Approach to AI Fairness-by-Design
ML	Machine learning
NLP	Natural Language Processing

Contents

Introduction

The AEQUITAS project provides a robust technical tool that detects and mitigates bias in Artificial Intelligence (AI). It also provides recommendations and methodologies that guide developers in achieving fairness in the early design stages. AEQUITAS **'Fairness-by-Design Engine' (FDE)** is developed as part of Work Package 5 (WP5) and is supplemented by the fairness-by-design methodologies created within Work Package 6 (WP6). The WP6 methodologies (also known as 'building blocks') identify relevant elements that can be applied throughout the AI lifecycle to facilitate fair AI innovation. Also known as '**building blocks**', the WP6 methodologies identify relevant elements that can be applied throughout the AI lifecycle, facilitating fair AI innovation. This document addresses one of those building blocks, namely, the **'Fair Data Collection, Governance and Management Methodology' (FDCGM)**.

The aim of the **FDCGM** is to aid **Development Decision Makers (DDM)** in ensuring their data collection, governance and management methods align with **ethical**, **legal** and **social** principles. The methodology integrates multiple dimensions of fairness: social, ethical, legal, and technical. This holistic approach recognizes that achieving truly fair AI systems requires careful consideration of data collection and management practices across all these perspectives. The FDCGM provides structured guidance through various stages of the AI development process, from initial scoping and risk assessment to data collection and evaluation.

This paper presents a detailed examination of data characteristics, types, and organizational frameworks relevant to AI development, while outlining specific methodologies for assessing and ensuring fairness in data practices. Furthermore, it offers practical guidance for implementing these methodologies by taking DDMs through a **Data Analysis Exercise.**

The methodology presented here represents a significant step forward in operationalizing fairness-by-design principles in AI development, providing developers with concrete tools and frameworks to address potential biases and ensure equitable outcomes from the outset of system development.

Fairness by Design

The Fair Data Collection, Governance and Management Methodology (FDCGM) serves as a step-by-step guide for AI developers to achieve **fairness-by-design** when it comes to data-related activities that impact subsequent development, outcomes, and performance throughout the entire AI lifecycle.

Figure 1: AI Lifecycle

Achieving fairness-by-design necessitates embedding techniques and practices into technological systems from the outset, ensuring their inherent fairness across the previously stated perspectives throughout all development phases. Therefore, '**fair-by-design'** can be defined as a proactive approach to developing technology in a manner that goes beyond harm mitigation after the placement of the system into the market and embeds fairness principles from the very beginning and throughout the technology lifecycle.

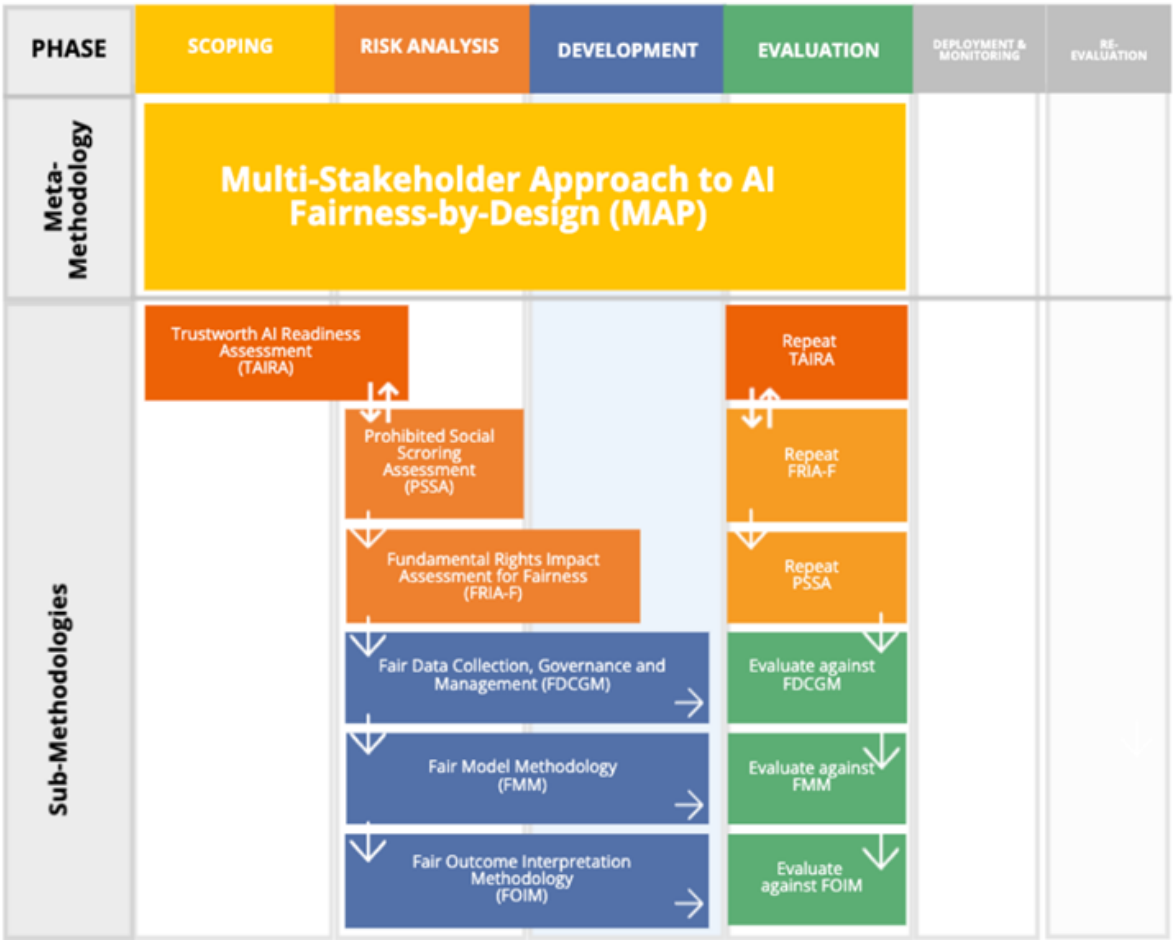

Figure 2: Fair-by-design methodology

Data collection, governance and management (CGM) represents a foundational cornerstone of AI system development. Within the AI lifecycle, CGM is particularly crucial during three key phases: scoping, risk assessment, and the data collection step of development (See Figure 2).

The scoping and risk assessment phases are critical for preventing bias in AI systems and unfair outcomes. During these phases, DDMs define the AI system's purpose and intended effects, determine appropriate data requirements including relevant features, variables, and labels, and establish suitable CGM methods. These early decisions are essential to fairness-by-design principles. By clearly specifying the system's purpose and comprehensively assessing potential risks during these initial phases, DDMs create the foundation for ethical and lawful AI development from the outset.

Notably, research conducted within AEQUITAS' WP6 has identified that 'fairness' manifests in four distinct but interrelated conceptions: **Social Fairness**, **Ethical Fairness**, **Legal Fairness** and **Technical Fairness**. These perspectives must be considered holistically throughout the CGM process to achieve fair AI systems.

Social Fairness

Social fairness emphasizes equality of opportunity and depends heavily on meaningful stakeholder engagement. This perspective considers how different groups may be affected by data collection and use, focusing particularly on historically marginalized or vulnerable populations. It involves ensuring equal representation in datasets and understanding the social context in which data is collected and used.

Ethical Fairness

Ethical fairness reflects how fairness manifests in relation to established AI ethics frameworks, particularly the EU's Ethics Guidelines for Trustworthy AI. This includes considerations of human agency, privacy, transparency, and accountability in data practices. The ethical perspective emphasizes the moral implications of data collection and use beyond mere technical or legal compliance.

Legal Fairness

Legal fairness centres on compliance with regulatory frameworks and the protection of fundamental rights. This includes adherence to the EU AI Act's data quality requirements, GDPR principles for data processing, and protections against discrimination. Legal fairness provides concrete requirements that shape data collection and management practices while ensuring human rights are protected

Technical Fairness

Technical fairness focuses on statistical parity and algorithmic bias mitigation. This perspective involves quantitative measures of fairness in data representation, processing techniques, and outcomes. It includes considerations of dataset balance, feature selection, and statistical bias detection methods, aiming to ensure fair treatment across different groups in technical terms.

Data and Data Sets

Before exploring how to ensure fairness in data collection, governance, and management practices, it is essential to understand what constitutes data in AI systems. This section provides an overview of fundamental data concepts, characteristics, and classifications that are relevant for AI development. It introduces readers to different types of data and datasets, explaining how they are defined both technically and within the EU regulatory framework. While later sections will address how to evaluate and ensure fairness in data practices, this section focuses on establishing a shared understanding of key terminology and concepts that form the basis for fair AI development.

The section is structured in two parts: first, examining the fundamental characteristics and types of data. Second, exploring how data is organized into datasets. This foundational knowledge is important for developers who aim to implement fair data practices throughout the AI lifecycle.

Data Characteristics and Types

Data is the accumulation of raw information that can be processed or transmitted to be rendered meaningful. The EU's Data Act defines data broadly as "any digital representation of acts, facts, or information and any compilations of such acts, facts or information, including in the form of sound, visual, or audiovisual recording[1]."

Levels of Representation

Though data is used as an umbrella and all-encompassing term, it is also important to distinguish between the levels of representation in data.

- Variable/Feature: As building blocks for data-based AI systems, features are those individual, measurable characteristics that are extracted from raw data. The process wherein the selection and preparation of features occurs is known as 'feature engineering'[2]. If we imagine our data to be structured in a tabular format, where each row corresponds to a set of measurements pertaining to a specific entity (e.g., a candidate during a hiring process, or a time stamp in a time-series), the variables/features would be the columns of such a table.

Data Point: Single strands of information that, compiled with similar strands could compose a feature. Each data point is assigned a value that is, usually, based on commonly agreed upon scales.[3] Each data point can be seen as a "row" if we adopt the tabular dataset format; a data point can comprise multiple features (that is, the entity that is the kernel of the data might be described by multiple variables).**Fundamental Characteristics**

When assessing data for AI systems, five key characteristics are considered:

- Volume: The amount of data, which impacts the scalability of AI outputs
- Velocity: The speed at which data arrives, particularly important for real-time analysis
- Variety: The range of data types and sources, requiring flexible AI models
- Veracity: The truthfulness and reliability of data
- Value: The significance of data in achieving the AI system's objectives

Data Organization Forms

Data can be organized into three main structures[4]:

- Structured Data: Neatly organized (e.g., in tables), commonly used in industrial applications
- Unstructured Data: Raw formats like text corpora or image sets, used for more versatile AI systems

[1] Regulation (EU) 2023/1854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act), Article 2(1).

[2] For an explanation of feature engineering see: What is a feature engineering? | IBM What is Feature Engineering? - Feature Engineering Explained - AWS (amazon.com)

[3] See for example: Data Points - What They Are and How to Use Them | Lenovo US

[4] See for example: Structured vs. Unstructured Data: What's the Difference? - IBM Blog Structured, Semi-Structured, and Unstructured Data | by RaviTeja@0401 | Medium Different data types: structured, semi-structured, and unstructured.... | Download Scientific Diagram (researchgate.net)

- Semi-structured Data: A hybrid form with some organizational features but less rigid than structured data

Data Types

Data types are categories that permit the classification of data based on its characteristics, organization, usage, format, limitations, manipulation, and analysis. In the context of AI, understanding the different types of data is a crucial step for CGM. Though many data types can be considered, we will present only some categories that are relevant to the EU legal field, AI systems and ensuring fairness in AI development and application.

Real-Life Data	Data that stems from and already exists in the world.
Personal Data	Under the EU's General Data Protection Regulation (GDPR), **personal data** refers to "any information relating to an identified or identifiable natural person[5]". The natural person in question is referred to as **data subject**, becoming so when their persona or identity can be distinguished from that of others regardless of the information that leads to their identification[6], i.e., numeric, license plates, name, pictures, etc. According to the Working Party 29's[7] opinion on personal data, several elements of this definition need to be retained, namely: ● any information, regardless of its nature, content or format can constitute personal data[8] ● the information itself 'relates to' a person if it provides direct or indirect content about the concerned person (e.g., identity, characteristics, behaviour, etc.) or if the person can be purposely evaluated on the basis of that information, or the if the information can influence the outcome of a decision pertaining to the identified person[9] ● while an identified person refers to the person that can be observed separately from the rest of the group on the basis of reasonable means to conduct identification, an identifiable person implies that identification through reasonable means is possible despite it yet not having occurred.[10]

[5] GDPR, Article 4(1), Recitals 14, 15, 27, 27 and 30. See also: What is personal data? - European Commission (europa.eu), WP29. Opinion 4/2007 on the concept of personal data.

[6] GDPR, Article 4(1)

[7] Article 29 Working Party (WP29) was the institution established under Article 29 of the EU's 1995 Data Protection Directive. Until its dissolution, WP29 was an independent advisory body established in relation to the European Union that provided advice on data protection and privacy. This institution was replaced by the European Data Protection Board in 2018 as a result of the GDPR's coming into effect. See: Legacy: Art. 29 Working Party | European Data Protection Board (europa.eu) Article 29 Working Party (iapp.org) Legal Framework | European Data Protection Board (europa.eu)

[8] 12251/03/EN (europa.eu) pp. 6-9

[9] 12251/03/EN (europa.eu) pp. 9-11

[10] 12251/03/EN (europa.eu) pp. 12-17

	The definition and understanding of personal data as prescribed in the GDPR and interpreted by the WP29 in relation to the Data Protection Directive[11], is employed across the EU landscape of digital law (e.g., Data Act, Article 2(3); Artificial Intelligence Act, Article 3(50)).
Non-personal data	According to the AI Act (Article 3(51)) and the Data Act (Article 2(4)), non-personal data concerns all data which cannot be classified as personal data in accordance with the GDPR's definition. In explanatory documents and court decisions, the European Union clarifies that non-personal data is data which has been transformed in such a way that it can no longer fulfil the criteria of personal data or allow the re-identification of the natural person[12], and that which has never fulfilled the same. In other words, if processes such as anonymisation and pseudonymisation can be reversed in a manner that permits identification, then the processed data remains personal data[13].

[11] In the Data Protection Directive, the definition of personal data can be found under Article 2(a). The Data Protection Directive has been repealed by the GDPR, and the definition has been adapted to include additional explanations. Nevertheless, the elements explained above can be identified in both the GDPR and the repealed Data Protection Directive.

[12] What is personal data? - European Commission (europa.eu) Storing and processing data in Europe: the free flow of non-personal data - Your Europe (europa.eu) See also: *SRB v EDPS*

[13] See: *SRB v EDPS* which rules that pseudonymized data is not personal data if it does not permit re-identification of the data subject

Special categories of personal data	Special consideration and protection are given to the types of personal data whose processing may lead to adverse effects or significant risks to fundamental rights, including for example discriminatory outcomes[14]. The widely coined terms **sensitive data** or **special categories of personal data** are generally used to describe this type of data[15]. Under the provisions of the GDPR, special categories of personal data include personal data that **reveals** racial or ethnic origins, political, religious and philosophical beliefs and opinions, an individual's trade union membership, or data that can be either genetic data, health data, and data concerning their sex life or orientation[16]. There are two important elements to retain about the GDPR's provision on special categories of personal data. Firstly, data that reveals certain special categories of attributes as described above, without actually being the attribute itself (see further proxy data) can fall under 'special categories of personal data'. Secondly, the processing of such data is prohibited unless the related activities occur by means of one of the exceptions listed under Article 9(2) of the GDPR, e.g., consent, the data subject's vital interest, public interest, etc[17]. Though the GDPR provides the rules for data processing activities concerning personal data (and special categories thereof), the AI Act governs the placement on the market and putting into use of AI systems. The Act offers a number of rules on data and data governance[18] and refers to the GDPR's provision regarding special categories of personal data as the retained definition for this type of data[19]. Without prejudice to the GDPR, the AI Act offers a further possibility to use of special categories of personal data within the training, testing and validation datasets of high-risk AI systems. **If the AI system at hand is a high-risk system, Article 10(5) of the AI Act specifies that special categories of personal data can be used (1) in exceptional cases, (2) strictly for bias mitigation and correction[20]. and (3) if sufficient safeguards, as described in Article 10.5. of the AIA, are implemented.** This exemption can be interpreted as a furthering of the GDPR's Article 9.2(g)[21]. This Article suggests that processing of special categories of personal data may occur if it is necessary for substantial public interest and is accompanied by EU or Member State law that is proportionate, respects the rights to data protection and provides suitable and specific safeguards[22]. From this perspective, the identification of special categories of personal data, already from the data collection stages and within the employed (training, testing and validation) data sets of high-risk AI is crucial because such data can only be employed if it has undergone a process of de-personalisation or anonymisation that renders the identification of individuals impossible.

Proxy data	The term 'proxy data' is derived from the social sciences and statistics 'proxy variables', i.e., variables that were traditionally used when the subject of a prediction could not be directly measured.[23] **Proxy data** is data that can infer, predict, or replace information. For example, data that represent facial expressions, blood flow, vocal tones, use of emojis can be used, based on an agreed-upon scale, to derive emotions.[24] Proxy data can, therefore, stem from a variety of sources. The use of proxy data in AI systems can overcome data scarcity or limitation, provide the system with additional background information that cannot directly be observed and, in some cases, serve the expansion of the data set towards more representative comments. However, the use of proxy data in AI systems raises more significant challenges that have the potential to affect fundamental rights to extreme degrees. Because of their inference techniques and black-box character, it is difficult to predict what attribute will be associated by AI systems to a certain proxy. It is likely, that an AI system will use proxies to derive attributes in a manner that voids human empathy and reasoning. For example, imagine an AI system that infers someone's nationality from the proxy data of 'language'. This system can easily make mistakes. For instance, someone might struggle to speak their native language fluently, but that doesn't mean they aren't a citizen of the country where that language is spoken. Likewise, someone who became a citizen after learning the country's language as a third or fourth language might not speak it perfectly, but that doesn't mean they aren't truly citizens[25].
Synthetic, stimulated, or augmented data	Deriving from traditional forms of data and data collection methods, AI systems are also capable of generating data. In other words, AI-produced synthetic or augmented data is data which has been created through digital means and which may serve to replace or mirror real-life data, protect some categories of data from being processed and improve the AI system[26]. Accordingly, organisations may opt to employ synthetic data either in the absence of real-life data or because the latter may be prone to biases.[27] Synthetic data is derived from data augmentation techniques, i.e., techniques that serve the artificial generation and increase of existing data points.[28]

[23] Warner, R., Sloan, R.H., *Making Artificial Intelligence Transparent: Fairness and the Problem of Proxy Variables.* Criminal Justice Ethics. (2021). 4-6.
[24] The Ethics of Emotion in AI Systems final submission (acm.org) p. 784
[25] For a real-life example of the negative effects of proxy data see: Inside the Suspicion Machine | WIRED

[26] What is synthetic data? | IBM Research Blog What is synthetic data? - MOSTLY AI
[27] Is Synthetic Data the Future of AI? (gartner.com)
[28] What is Data Augmentation? Techniques & Examples in 2024 (aimultiple.com)

Biometric data	The definition of biometric data can be firstly derived from Article 4(14) of the GDPR which refers to personal data (see above) that results from specific technical processing that relates to the following characteristics of the natural person: (1) physical, (2) physiological and (3) behavioural. The data that can relate to these characteristics may also include but is not limited to facial images or dactyloscopic data. In line with the definition of personal data, the GDPR further notes that the specific technical processing should either permit or confirm the natural person's unique identification[29]. In line with this description, but without holding a direct reference to the GDPR, the AI Act defines biometric data as: *"(…) personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person; such as facial images or dactyloscopic data;"*[30]
Data concerning health	Defined in the GDPR, data concerning health is that data which relates to or reveals the mental or physical health as well as the health status of an individual. The use of the terms 'related to' and 'reveals' implies that data which indirectly allows the recognition of any health-related matters concerning the natural person (see also: proxy data) would fall in this category[31]. 'Data concerning health' is considered a special category of personal data (see above)[32].
Metadata	Metadata refers to the aggregated information that describes a cluster of data by outlining, among others, relevant characteristics, relations among data points, and technical details[33]. In EU law, the Data Act considers metadata to be *"a structured description of the contents or the use of data facilitating the discovery or use of that data"*[34]. Understanding the metadata of a given dataset permits the identification and analysis of different errors as well as their sources whilst also allowing the informed application of e.g., bias-mitigation techniques.
Readily available data & primary data	Akin to what scientists refer to as **secondary data,** readily available data can be considered in opposition to **primary data**. The latter is data which has been collected first-hand by an institution, group or natural person that significantly decided upon the collection techniques (e.g., survey, experiments, etc), data variables, data types, and original purpose for the data. Subsequently, **readily available/ secondary data** is that which has been collected by another entity than the one which will be using the data, and which may be composed of e.g., historic data, records, openly available datasets, etc[35]. In EU law, the term '**readily-available data'** does take a slightly different meaning, describing that data contained in a product or a service, that an institution has or can lawfully obtain from said product/service without undergoing disproportionate efforts (e.g., a simple operation)[36]. Combined with the above explanation, the secondary/different data collectors would be the product or service in question, whereas the entity retrieving that data would be the one using it.

[29] See Article 4(14), GDPR.
[30] Article 3(34), AI Act.
[31] Article 4(15), GDPR
[32] Article 9(1), GDPR.
[33] Metadata: What it is and how it works | NordVPN Definition - Metadata for Data Management: A Tutorial - LibGuides at University of North Carolina at Chapel Hill (unc.edu)
[34] Article 2(1), Data Act.
[35] See for example: Primary & Secondary Data Definitions - Public Health Research Guide - Research Guides at Benedictine University Library Primary vs. Secondary Data | Explanation, Uses & Tips - ATLAS.ti (atlasti.com)
[36] Article 2(1), Data Act.

Training Data	Data which is used for training of AI systems, i.e., that portion of the data which is given to the system on the basis of which the model learns and uncovers patterns[37]. Accordingly, the AI Act defines training data as *"data used for training an AI system through fitting its learnable parameters".*[38] In machine-learning, learnable or training parameters are settings that can be optimised to adjust the model's performance, in particular the weights or biases that are adjusted during the training process.
Input data	According to the AI Act, input data is that which an AI either acquires or is provided with, on the basis of which it can issue an output.[39]
Testing data	Data which is used for an independent evaluation that serves to confirm or assess an AI's expected performance before the system is either placed on the market or put into service (in the EU).[40]
Validation data	Data employed to assess the performance of an already trained AI system, serving to fine-tune the system's non-learnable parameters and learning process, therefore aiming to mitigate, among others, under-/overfitting.[41] Subsequently, the AI Act notes that the validation dataset is a dataset that is separate from the training dataset, either as a fixed or as a variable split.[42]

Datasets are organized collections of data that serve specific purposes in AI development. The EU's AI Act distinguishes three primary types:

Training Datasets

Training datasets are the first set of data used to train the model to understand patterns and relationships within the data from which it can then infer outcomes, predictions, decisions, etc. In some cases, a hold-out method is used in which the training dataset is split into 70/30, with 70 % of the data used as a training dataset and 30 % of the data as a validation dataset.[43]

Validation Datasets

The performance of AI is evaluated based on the discrepancy between predictions and actual values. Following this, validation datasets are employed to assess the accuracy of the model to ensure correct data classification and predictable results. This type of dataset is comprised of data that is new to the model and that is designed to test its performance.

---

[37] The Difference Between Training Data vs. Test Data in Machine Learning (obviously.ai)
[38] Article 3(29), AI Act.
[39] Article 3(33), AI Act.
[40] Article 3(32), AI Act.
[41] Article 3(30), AI Act.
[42] Article 3(31), AI Act.
[43] Devi, K. (2023, August 14). Understanding Hold-Out Methods for Training Machine Learning Models. Comet. https://www.comet.com/site/blog/understanding-hold-out-methods-for-training-machine-learning-models/

A common method of validation uses benchmark datasets to assess the quality of AI models. A benchmark dataset is a shared resource and presents a high-quality standard formulation of a specific machine-learning task with an accompanying quantitative metric of evaluation that other models can be compared to. [44]

Notably, (see also above) the validation dataset is defined under the AI Act as "a separate data set or part of the training data set, either as a fixed or variable split;"[45].

Testing Datasets

Testing datasets comprise entirely new data designed to assess the overall functioning of the model. This final evaluation phase uses data that the model has never encountered before, providing a crucial check of the model's ability to perform reliably in real-world conditions. Testing datasets serve as the ultimate verification of whether the AI system can generalize its learned patterns to novel situations and produce accurate outputs when deployed.

Data Collection, Governance and Management

Under the AI Act, the EU establishes several principles and requirements to be applied to data sets used in AI system development. To facilitate the practical implementation of these requirements while ensuring fairness across multiple dimensions, we have developed a comprehensive **Data Evaluation Exercise** that forms a critical component of the Fair Data Collection, Governance and Management Methodology (FDCGM), the template for which can be seen in Annex 1. This **Data Evaluation Exercise** is designed to address fairness through four interconnected dimensions: **legal**, **ethical**, **social**, and **technical**.

The **legal dimension** is directly incorporated through assessment criteria that explicitly map to **EU AI Act requirements**, providing Development Decision Makers (DDMs) with clear guidance for regulatory compliance. The **ethical dimension** builds upon work completed during the earlier **Trustworthy AI Readiness Assessment** stage, where systems were evaluated against the **EU's Ethics Guidelines for Trustworthy AI** requirements. DDMs are encouraged to reference and build upon these earlier ethical evaluations when conducting their data assessment.

The **social** and **technical** dimensions of fairness are addressed through structured evaluation columns in the exercise that prompt DDMs to identify potential gaps or shortcomings in terms of fairness, and to develop specific mitigation strategies aligned with the **EU AI Act**. These columns facilitate systematic consideration of both **social implications**, such as impacts on different demographic groups, and **technical considerations**, such as statistical bias and data quality issues.

First, we propose a **dataset-level assessment** examining eleven key aspects derived primarily from Article 10 of the EU AI Act, incorporating both compliance requirements and fairness considerations:

[44] Koch, B., Denton, E., Hanna, A., & Foster, J. G. (2021). Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research (arXiv:2112.01716). arXiv. https://doi.org/10.48550/arXiv.2112.01716

[45] Article 3(31), AI Act.

- Data characteristics (volume, velocity, variety, veracity, value)
- Data type classification (personal, non-personal, special categories, etc.)
- Error assessment and quality validation
- Data source evaluation
- Purpose of data collection
- Data collection methods assessment
- Data processing methods review
- General statistical robustness verification
- Representativeness evaluation
- Relevance assessment
- Data Fairness

This initial evaluation provides both a pathway toward partial legal compliance with the EU AI Act and a foundation for ensuring fairness in AI systems. By examining these aspects at the dataset level, we can identify systemic issues that might affect the overall fairness and compliance of the AI system.

Following this comprehensive dataset evaluation, the methodology proceeds to **assess each individual feature within the dataset**, examining specific fairness implications and compliance requirements at a granular level. This two-stage approach ensures thorough consideration of both system-wide and feature-specific fairness concerns throughout the data collection, governance, and management processes.

Lastly, DDMs are required to provide technical documentation of the data evaluation process and outline clear fair data governance strategies, which will be presented in the form of a **Data Dictionary** at the end of the Data Evaluation Exercise.

Data Governance and Management

The Fair Data Collection, Governance and Management Methodology (FDCGM) is fundamentally anchored in the regulatory requirements set forth by the **EU AI Act**. These requirements form the essential framework through which the **Data Evaluation Exercise** assesses datasets and their governance structures. Before proceeding with specific evaluation methodologies, it is crucial to establish a clear understanding of these **legal concepts** that will guide the assessment process. This section therefore examines the key regulatory requirements that Development Decision Makers (DDMs) must consider when implementing data governance frameworks that aim to achieve both legal compliance and fairness.

Relevant design choices	The AI Act notes that data governance practices of (high-risk) AI systems should include design choices within their training, testing and validation data sets, that are relevant to the intended purpose of the AI system. In other words, data should be collected and processed in a manner that is relevant to the AI system's intended purpose. Ensuring the relevancy of all design choices requires the identification and evaluation thereof.

Data collection processes and the origin of data, and in the case of personal data, the original purpose of the data collection	The EU AI Act requires organizations to carefully track and document their data collection processes and where their data comes from. For data collection, organizations need to show exactly how they gathered the data and prove that their collection methods make sense for what their AI system is trying to accomplish. When it comes to data origins, organizations must be able to identify and verify all their data sources. This means keeping clear records of where each piece of data came from and confirming those sources are reliable. If the data was collected from somewhere else rather than gathered directly, organizations need to understand and document the original context it was collected. Personal data requires extra attention. Whenever organizations collect or use personal data, they must document what it was originally collected for and make sure they're following data protection laws like GDPR. They need to have a valid legal reason for using the personal data and ensure they're only using it in ways that match its original purpose.
Relevant data-preparation processing operations, such as annotation, labelling, cleaning, updating, enrichment and aggregation	The collected data must be subject to adequate data processing activities, including, for example, labelling, cleaning, enrichment and aggregation. The relevant data processing operations should be identified, clearly documented and evaluated for their effect on the AI system's fairness. For instance, aggregation is a notorious source of bias, therefore the performance thereof should be clearly stated.
The formulation of assumptions, with respect to the information that the data are supposed to measure and represent	Think of this requirement as stating a research hypothesis and elaborating how you expect the collected data will help test this hypothesis. For instance, suppose you are designing an AI tool that predicts traffic flows of a city to reduce traffic jams, and you use a (secondary) traffic collection dataset collected through inductive loops. Here, the assumption is that the collected data will share insight into the busiest roads of a city, which will allow you to better predict where jams occur.

An assessment of the availability, quantity and suitability of the data sets that are needed	Organizations need to carefully evaluate three key aspects of their datasets: whether they can actually access the data they need, if they have enough of it, and if it's right for their purpose. For availability, organizations must confirm they have proper access to all necessary data and have appropriate permissions or agreements in place to use it. This includes checking any licensing requirements or data sharing agreements. When assessing quantity, organizations need to determine if they have enough data to train their AI system effectively. This means having sufficient data for training, testing, and validation purposes. The amount needed will vary based on the complexity of the AI system and its intended use. Suitability means making sure the data is appropriate for the AI system's purpose. Organizations should evaluate whether their datasets are relevant, representative of the real-world scenarios the AI will encounter, and of high enough quality. This includes checking that the data is accurate, up to date and contains the right kinds of information needed for the system to function properly.
Examination in view of possible biases that are likely to affect the health and safety of persons, have a negative impact on fundamental rights or lead to discrimination prohibited under Union law, especially where data outputs influence inputs for future operations	Organizations must examine their data for potential biases that could cause harm in three main ways: affecting people's health and safety, impacting fundamental rights, or leading to prohibited discrimination. This examination needs to happen at every stage of data handling - from collection through processing to use - and should identify both obvious and subtle forms of bias. Organizations must document any biases they find, assess who might be affected, and create specific plans to remove or reduce these biases.
Appropriate measures to detect, prevent and mitigate possible biases identified	Organizations must implement specific measures to handle biases in their data. This means having clear procedures to detect biases through regular testing and monitoring, preventing new biases from emerging through careful data collection and processing practices, and mitigate any biases that are found. These measures should be documented and regularly reviewed to ensure they're working effectively. Each measure should address specific identified biases with concrete steps, not just general principles. The effectiveness of these measures should be tested and verified, with adjustments made when needed. If new biases are discovered, organizations should quickly assess them and add appropriate countermeasures to their existing bias mitigation strategy.

The identification of relevant data gaps or shortcomings that prevent compliance with this Regulation, and how those gaps and shortcomings can be addressed.	Organizations must actively identify any gaps or weaknesses in their datasets that could prevent them from meeting legal requirements. This involves checking for missing data, incomplete information, or quality issues that might affect compliance. When gaps are found, organizations need to create specific plans to address them. Special attention should be paid to gaps that might affect fundamental rights or lead to discrimination.

Building upon the foundational principles and practices discussed earlier in this section, we must delve deeper into the intricate elements that collectively shape the responsible management and governance of data within AI systems. These elements uphold the integrity and quality of data but also serve as guardians of fairness. They ensure that the outputs and decisions rendered by these systems align with ethical standards and mitigate the perpetuation of harmful biases.

Appropriate performance metrics	Fairness metrics are quantitative measures that assess the performance of an AI system across various demographic groups or protected classes. These measures assist in identifying potential biases or inequalities in the model's predictions or outputs. Fairness metrics include statistical parity, equal opportunity, and differential impact. Assessing suitable fairness metrics is critical because it enables developers to identify and address unfair treatment or discrimination against specific groups. By measuring and monitoring these indicators, developers may verify that their AI systems make fair and equitable decisions without unexpected biases.
Test logs	Test logs are comprehensive records of the testing process for an AI system. They document the various tests conducted, including their inputs, expected outputs, and actual results. These logs are crucial for data management and governance, as they enhance transparency and accountability throughout the development process. Test logs enable developers to track and analyze the performance of their AI systems, identify potential issues or biases, and make informed decisions regarding improvements or corrective actions. Furthermore, test logs serve as evidence of the measures implemented to ensure fairness and compliance with relevant regulations and ethical guidelines.
Test reports	Test reports encapsulate the findings and insights derived from the testing process. These reports typically encompass an overview of the testing methodology, detailed results from various tests (including fairness evaluations), and recommendations for further improvements or mitigations. Test reports provide a structured mechanism to communicate testing outcomes to stakeholders, including developers,

	decision-makers, and auditors. By meticulously documenting the fairness and performance of the AI system, test reports enhance transparency, accountability, and informed decision-making concerning the deployment and utilization of the AI system.
Labelling Procedures	Labelling procedures are the processes and guidelines for annotating or labelling the data needed to train AI models. Proper labelling processes are critical to guaranteeing the quality and fairness of training data. This includes developing clear and explicit labelling criteria, forming diverse and representative labelling teams, and adopting quality control techniques to reduce errors and biases. Well-defined labelling techniques help limit any biases in the training data, which might lead to unfair or discriminating results in the AI system's predictions or choices.
Data Cleaning	Data cleaning is the process of finding, correcting, or deleting errors, inconsistencies, or extraneous information from the data needed to train artificial intelligence models. This procedure comprises processing missing values, deleting duplicates, and finding and treating outliers or anomalies. Data cleaning is critical for data management and governance because it enhances the quality and reliability of training data, which has a direct impact on the fairness and accuracy of the AI system's results. By cleaning and preparing the data, developers may minimise potential biases or distortions produced by low-quality or inconsistent data, ensuring that the AI system is trained on a clean and representative dataset.

Data Source Selection and Collection Methodologies

To proactively produce a fair-by-design system, the sources of the employed data must reflect fairness principles. Broadly, AI systems may be trained with data stemming from two types of data sources that have previously been categorised as either primary or secondary.

Primary Data Sources

Primary data sets are comprised of (1) data acquired directly from the source and (2) with the entity collecting the data being the one that utilizes the data, or if it is an agency that collects the data, it gathers it on this entity's behalf. The key advantage of adopting primary data sources and collection methods is that you have more control over the data. This control is especially useful if you have previously determined your data requirements, as in the preceding steps. In such circumstances, primary data provides the greatest flexibility to match your individual needs and desired scale. However, creating a primary dataset large enough for AI can be resource-intensive and time-consuming, making it unsuitable for individuals without the necessary resources.

Several methods can be employed for primary data collection, with the choice depending on the specific research objectives. These methods include:

Surveys and Questionnaires	Surveys and questionnaires are tools that serve to gather information from a defined and specific. They require a process of question construction that is purpose and context-specific.
	
	To proceed with this method, it is important to ensure that the questions are formulated to induce the least amount of biased responses, and that the population parameters, as well as the limitations thereof are clearly stated. To achieve fairness in surveys, consider the following questions:
	
	• **Social Perspective**:
	  ○ Does the survey address potential disproportionate impacts?
	  ○ Does the survey take into account group differences?
	  ○ Which safeguards are implemented to prevent discrimination?
	  ○ Which demographic, contextual or geographic limitations have been observed?
	  ○ Have the relevant stakeholders assessed the survey for its social fairness?
	• **Ethical Perspective:**
	  ○ Does the survey respect human values?
	  ○ Is there sufficient documentation over the survey's development?
	  ○ Have the relevant stakeholders assessed the survey for its ethical fairness?
	• **Legal Perspective:**
	Consider whether special categories of data will be collected and comply with relevant legislation.
	Ensure the transparency and correctness of information in the survey.
	If any balancing of fundamental rights occurred, ensure transparency and clarity thereover.
	Ensure clear accountability paths regarding the production and distribution of the survey, and subsequent data collection.
	• **Technical perspectives:**
	  o How will potential biases be technically mitigated?
	  o How will the survey impact the AI system's capacities?

Observations	Observations involve the recording of patterns resulting from the process of carefully observing individuals or events exactly as they occur, undisturbed, in real settings. Non-participant and unstructured observations are examples of different types of observations. Some of these forms involve the construction of protocols and require the placement of the data into machine-readable formats.
	
	To render observations fair, consider the following:
	• **Social Perspective:**
	  ○ Do the observations capture interactions and outcomes in manners that permit an equal assessment of stakeholders?
	  ○ What safeguards are set in place to mitigate discriminatory data collection?
	• **Ethical Perspective:**
	  ○ How were ethical considerations respected during the observation process?

	• **Legal Perspective:**
	  ○ Consider whether special categories of data will be collected and comply with relevant legislation.
	  ○ Ensure the transparency and correctness of information in the survey.
	  ○ If any balancing of fundamental rights occurred, ensure transparency and clarity thereover.
	  ○ Ensure clear accountability paths regarding the production and distribution of the survey, and subsequent data collection.
	• **Technical Perspective:**
	  ○ How will potential biases be technically mitigated?
	  ○ How will the survey impact the AI system's capacities?

Sensors and Devices	Data collection that makes use of either sensors or electronic devices. To lead this form of data collection, which is more prevalent within the context of AI system development than other forms of primary data collection, it is important to consider the following steps:
	
	• Identify the type of sensor (motion, biometric, image, location, etc.) or electronic device (e.g., smartphone, watch, wearable) that is most appropriate given the intended purpose.
	• In the case of the sensor, ensure its strategic placement.
	• Derive and establish the level of consent and user awareness of data collection given the intended purpose and data CGM protocol.
	• Conclude a (data protection) impact/risk assessment and implement fair risk mitigation measures.
	• Monitor the collected data and the desired parameters. Ensure that no additional parameters are being measured.
	• Store and retain data correctly and appropriately storage of data.
	• Apply appropriate data processing methods including, e.g., filtering, cleaning, aggregation, pseudonymisation and anonymisation.
	
	The questions attributed to different forms of fairness should be considered and complemented during and throughout each of the previous steps. Other method-specific questions that enable fairness include:
	• **Social Perspectives:**
	  ○ How do the sensors and devices react to different demographic characteristics?
	  ○ In the case of sensors combined with the collection of biometric data, is there sufficient consideration of differences among physical attributes?
	  ○ What are the safeguards implemented to mitigate disparities in incorrect measurements?
	• **Ethical Perspectives:**
	  ○ How were ethical considerations identified in the **Trustworthy AI Readiness Assessment** and the **Trustworthy AI Leaflet** respected during the observation process?
	• **Legal Perspectives:**
	  ○ Does the use of sensors intervene with any fundamental right (e.g., data privacy)? If so, which safeguards are implemented to mitigate

	potential harm? Consider leading AEQUITAS' Fundamental Rights Impact Assessment Methodology (FRIA). ○ Consider whether special categories of data will be collected and comply with relevant legislation. ○ Ensure the transparency and correctness of information in the survey. ○ If any balancing of fundamental rights occurs, ensure transparency and clarity thereover. ○ Ensure clear accountability paths regarding the production and distribution of the survey, and subsequent data collection. • **Technical Perspectives:** ○ What protocols were implemented to ensure the consistency of the sensor / device-based data collection? ○ What protocols were implemented to ensure the accuracy of the sensor / device-based data collection? ○ What protocols were implemented to ensure the representativeness of the sensor / device-based data collection? ○ What protocols were implemented to ensure the reliability of the sensor / device-based data collection? ○ In the case of sensors, how were the sensors validated?
Social Media and Online Fora	Data collection that occurs through the medium of social media or other online fora. Consider the following process: • Identification of relevant platforms and fora that respond to the intended purpose. • Select data collection methods, e.g., internet scraping, surveys, automated content analysis, network analysis, etc. The chosen method should be clearly analysed from the different perspectives of fairness. • Consider, ethical considerations. Due to the ubiquitous form of data collection, it is important to ensure that rights are not infringed and that concerns related to consent and data ownership are resolved. Adhere to the **EGTAI**, ALTA and **GDPR.** • Apply appropriate data processing methods e.g. cleaning, aggregation, pseudonymisation and anonymisation. To ensure fairness, consider answering the following questions: • **Social Perspective:** ○ How does data collection from social media or online fora propagate discrimination and bias? ○ Are there any groups or segments of the social media platform or online forum that engender discriminatory or biased results in data? • **Ethical Perspective:** ○ How were ethical considerations respected during the observation process? You can refer back to the **Trustworthy AI Readiness Assessment** and the **Trustworthy AI Leaflet** • **Legal Perspective:** ○ Does the use of social media and online fora intervene with any fundamental right (e.g., data privacy)? If so, which safeguards are implemented to mitigate potential harm? Consider leading AEQUITAS' Fundamental Rights Impact Assessment Methodology (FRIA). ○ Consider whether special categories of data will be collected and comply with relevant legislation.

	○ Ensure the transparency and correctness of information in the survey. ○ If any balancing of fundamental rights occurs, ensure transparency and clarity thereover. ○ Ensure clear accountability paths regarding the production and distribution of the survey, and subsequent data collection. • **Technical Perspectives:** ○ What protocols were implemented to ensure the consistency of the social media/ online forum data collection? ○ What protocols were implemented to ensure the accuracy of the social media/ online forum data collection? ○ What protocols were implemented to ensure the representativeness of the social media/ online forum data collection? ○ What protocols were implemented to ensure the reliability of the social media/ online forum data collection?
Transactional Data Tracking	The collection of data pertaining to interactions amongst human or institutional entities. The contained information may be representing purchases, interactions with online fora, digital exchanges, financial transactions, etc. This data may be retrieved from point-of-sale systems, mobile apps, or other digital channels. Where this data includes information about customers, that can be considered personal data, and, depending on the uptaken details, special categories of personal data. This form of data collection is typically used for entrepreneurial insights, fraud detection, inventory management, economic monitoring, and compliance. To ensure fairness in transactional data tracking, consider answering the following questions: • **Social Perspective:** ○ How does data collection from transactional data tracking methods propagate discrimination and bias? ○ Are there any groups or segments of transactional data tracking methods that engender discriminatory or biased results in data? • **Ethical Perspective:** ○ How were ethical considerations respected during the transactional data tracking methods process? • **Legal Perspective:** ○ Does the use of transactional data tracking methods intervene with any fundamental right (e.g., data privacy)? If so, which safeguards are implemented to mitigate potential harm? Consider leading AEQUITAS' Fundamental Rights Impact Assessment Methodology (FRIA). ○ Consider whether special categories of data will be collected and comply with relevant legislation. ○ Ensure the transparency and correctness of information in the survey. ○ If any balancing of fundamental rights occurs, ensure transparency and clarity thereover. ○ Ensure clear accountability paths regarding the production and distribution of the survey, and subsequent data collection. • **Technical Perspectives:** ○ What protocols were implemented to ensure the consistency of the transactional data tracking methods?

	○ What protocols were implemented to ensure the accuracy of the transactional data tracking methods?

Secondary Data Sources

Secondary datasets are collections of data that were originally gathered by entities other than the organization currently seeking to use them. When an organization wants to use these datasets in AI systems, it must ensure the data meets specific standards of integrity, fairness, and compliance with relevant legislation such as the GDPR and AI Act.

These datasets must fulfil numerous requirements, including relevance, representativeness, accuracy, and completeness. They must also account for contextual considerations such as geographical, behavioural, and functional aspects specific to the AI system's intended use. These requirements will be thoroughly assessed in the Data Evaluation Exercise and will be further detailed in the next sections of this document.

Meeting these requirements presents unique challenges when working with secondary datasets, as opposed to primary data collection. Organizations often face difficulties in tracing the origin of the data or understanding the precise methodologies used in its collection. Furthermore, the original data collectors may not have followed the same rigorous standards now required for AI system development, or they may have had different objectives that affect the data's suitability. Privacy considerations, consent documentation, and bias monitoring can be particularly challenging when working with data collected by others.

To address these challenges, organizations can implement several strategies. First, they should establish direct communication with the original data collectors whenever possible to obtain comprehensive documentation about collection methods and data characteristics.

Organizations should also conduct thorough bias audits and implement robust evaluation frameworks to assess the data's suitability. Engaging diverse stakeholders in the review process can help identify potential issues and ensure the data's appropriateness for its intended use. When gaps or issues are identified, organizations should consider supplementing the secondary data with additional primary data collection or implementing specific processing techniques to address identified shortcomings. Maintaining detailed documentation of all evaluation steps and decisions made regarding the secondary dataset is crucial for ensuring transparency and accountability in the AI system's development process

Evaluating a dataset for fairness

Data Level Assessment

The first step in evaluating your data set, is assessing it at a data set level. This provides DDMs with an overarching view of all of the key characteristics.

The following section provides an explanation of each element that is assessed at the **data level** in the **Data Evaluation Exercise**. A template for the Data Evaluation Exercise can be seen in Attachment 1.

For each aspect, we provide specific evaluation criteria and considerations that span all four fairness dimensions, ensuring a holistic approach to data assessment. The **legal dimension** is addressed by assessing each element in accordance with the relevant parts of the **EU AI Act**. Whilst the **ethical dimension** was largely addressed in the **Trustworthy AI Readiness Assessment** stage, we provide some additional relevant considerations. The **social** and **technical** perspectives are also holistically integrated throughout the questionnaire.

The framework includes both compliance requirements and fairness considerations, providing DDMs with practical guidance for implementing fair data practices throughout the AI lifecycle.

Assessment Element	Description	Impact on Fairness	EU AI Act Reference
Data Characteristics	Examination of volume (amount of data), velocity (speed of data arrival), variety (range of data types/sources), veracity (truthfulness/reliability), and value (significance for objectives), sample size, number of data features, data transformation, data availability	When evaluating datasets, it is essential to consider how the data characteristics might affect fairness. When data volume, variety, or veracity is insufficient for certain groups, it can create a cascade of fairness problems. For example, if an AI system is trained on a dataset with 10,000 samples from one demographic but only 100 from another, it will likely perform worse for the underrepresented group. Similarly, if data velocity (speed of updates) varies between groups, some populations may be represented by outdated information while others benefit from current data. Poor veracity in data about minority groups can reinforce stereotypes or lead to incorrect assumptions being encoded into the system.	Article 10(2)(e) & Article 10(3)
Data Type Classification	Identification of the form in which the data is in (structured, unstructured), the data types present in the dataset (image, audio, etc.) and categorization of data into regulatory types (special categories per GDPR)	The handling of personal and special category data directly affects fairness because these categories often correlate with protected characteristics. Even when explicitly protected characteristics are removed,	Article 10(3), Article 10(4) & Article 10(5)

		seemingly neutral data can serve as proxies – for example, postal codes might correlate with racial demographics. Mishandling of special category data can lead to discrimination, while improper anonymization might expose vulnerable groups to risks.	
Error Assessment and Quality Validation	Examination of inconsistencies, duplicates, missing values, outliers, version consistency, measurement consistency, format consistency	Systematic errors in data quality can disproportionately affect certain groups. For instance, if missing data occurs more frequently for certain demographics, the system might make more assumptions or generalizations about these groups. Poor quality validation might allow biased or incorrect data to influence the system's decisions, particularly affecting groups that are already marginalized or underrepresented.	Article 10(3)
Data Source Evaluation	Assessment of data origin, repository, ownership, institutional affiliations, access rights, sharing agreements	The sources of data can embed historical biases and discriminatory practices. For example, if historical hiring data is used, it might reflect past discriminatory practices. Different data sources might have varied quality standards or collection methods for different groups, leading to systematic disadvantages in how certain populations are represented in the dataset.	Article 10(2)(b), Article 10(4) & Annex IV (1)(a)
Purpose of Data Collection	Examination of assumptions about what data measures/represents relative to system purpose		Article 10(2) & Annex IV (1)(a)

Data Collection Methods Assessment	Evaluation of collection processes, sampling methods, potential biases, inclusion/exclusion criteria	Collection methods can systematically exclude or under-represent certain groups. For example, online-only surveys might underrepresented elderly or low-income populations with limited internet access. The timing, location, or format of data collection can create barriers for certain groups, leading to skewed representation in the dataset.	Article 10(2)(b)
Data processing methods review	Examination of data preparation operations: annotation, labelling, cleaning, updating, enrichment, aggregation	Processing methods can introduce or amplify biases. For instance, standardization processes might normalize away important variations in how different groups express the same characteristics. Automated cleaning processes might disproportionately remove valid but unusual data points from minority groups, treating them as outliers. From a **technical fairness** perspective, it is thus important to consider how the processing methods used may affect various groups.	Article 10(2)(c), Article (10)(2)(f) & Article (10)(2)(g)
Statistical Assessment	Statistical properties are characteristics of the data such as distribution, means, and variability. These properties should be appropriate given the AI system's intended purpose and should, where appropriate, be considered in regards to the stakeholders with whom the AI system is intended to be used.[63] Assessment of statistical properties includes central tendency, variability, significance, etc.	Poor statistical robustness can lead to unreliable performance for underrepresented groups. When statistical properties aren't verified across all subgroups, the system might appear to perform well overall while failing for specific populations. This can mask significant disparities in system performance across different demographics	Article 10(3)

Representative	What is considered a data set that is sufficiently representative depends on the population that is observed in the data, its variability and the acceptable margin of error. Based on a thorough analysis of the involved stakeholders, phenomena and desired system output, the representative parameters may differ and so may what be considered sufficient.	Lack of representativeness directly affects system fairness by creating blind spots in how the AI system handles different groups. When certain populations are underrepresented, the system may make poor generalizations about these groups or fail to account for their specific characteristics and needs, leading to discriminatory outcomes.	Article 10(2)(f), Article 10(3) & Article 10(4)
Relevance Assessment	The data in the data set should be relevant to the AI's intended purpose or task that it is designed for. All elements which do not meaningfully and purposefully contribute to the task that the system is trying to fulfil should be removed.	Including irrelevant data can create spurious correlations that disadvantage certain groups. For example, if irrelevant socioeconomic indicators are included in a hiring system, they might create unwarranted barriers for certain populations. The system might learn to make decisions based on factors that have no real bearing on the task but correlate with protected characteristics.	Article 10(2)(d) Art. 5.1(c) Prohibited Social Scoring
Data Fairness	An assessment of how data collection, processing and usage might unfairly impact different groups through bias or discrimination. It requires identifying potential biases and implementing specific measures to detect, prevent and mitigate any unfair impacts.	Overall data fairness serves as a critical lens through which to examine potential discriminatory effects. When data fairness is compromised, it can lead to compound effects where multiple small biases combine to create significant discrimination. This can manifest in both direct discrimination (treating groups differently based on protected characteristics) and indirect discrimination (seemingly neutral practices that disadvantage certain groups).	Article 10(2)(f)

Data Feature Level Assessment

While **dataset-level assessment** provides a broad view of fairness considerations, individual features within datasets can introduce or amplify biases in ways that may not be apparent at the aggregate level. Thus, the second part of the **Data Evaluation Exercises** provides a framework for evaluating each data feature against **legal**, **ethical**, **social**, and **technical** fairness criteria.

This detailed examination is essential for identifying potential fairness issues that might emerge from specific data attributes or their interactions.

Beyond assessing the below elements, the **Data Evaluation Exercise** template seen in Attachment 1 also requires DDMs to evaluate whether their assessment below presents any gaps or shortcomings in terms of fairness. Additionally, DDMs are required to think through solutions on how these gaps can be mitigated and addressed, in compliance with the EU AI Act.

Assessment Element	Description	EU AI Act Reference
Type of Data Feature	Classification of features based on whether it is personal data under GDPR, special category data, or could serve as a proxy for protected characteristics. Includes determination of data format (structured/unstructured) and basic type (numeric, categorical, text, etc.)	Article 10(5)(b) & Article 10(5)(c)
Scale	The measurement scale used for the feature, including its range and units of measurement	Article 10(2)(f) & Article 10(3)
Volume	The amount of data available for this feature, considering both overall quantity and distribution across different groups	Article 10(2)(f) & Article 10(3)
Distribution	How values are distributed across the feature, including consideration of central tendency, variance, and outliers	Article 10(2)(f) & Article 10(3)
Relevance	How necessary and appropriate the feature is for the system's intended purpose	Article 10(2)(d) & Article 5.1(c)
Free of Errors	Assessment of data quality issues including missing values, inconsistencies, and measurement errors	Article 10(2)(f) & Article 10(3)
Complete	Whether the feature provides comprehensive coverage across all relevant populations and contexts	Article 10(3) & Article 10(4)

Data Feature Interaction	How the feature interacts with and relates to other features in the dataset	Article 10(2)(f) & Article 10(3)

Transparency and Documentation

In order to keep track of necessary changes within the data, the data sets and subsequently, the data collection process, it is advised to ensure that sufficient documentation occurs. Within the **Data Evaluation Exercise**, participants are required to fill out a **Data Dictionary.** The Data Dictionary provides a clear overview of all of the assessments that occurred at both Data and Feature Level Assessments.

However, the Dictionary is not the only form of documentation. The other forms of frameworks include:[95]

Data Cards/ Data Sheets	Summaries that provide information about the dataset that is used to train or evaluate AI models. It includes information relating to collection and processing activities. The result of undertaking the present methodology will be, on the one hand, ensuring fair data collection, and on the other, producing a data card/data dictionary that enables the transparent documentation thereof.[46] Notably, datasheets have been named in the AI Act as an appropriate form of disclosure for fulfilling requirements related to the presentation of information related to data.[47]
Model Cards	Similarly to data cards, model cards provide information about the AI model. It includes a description of the AI system's limitations, uses, performance capacities, and, at times, evaluated ethical considerations.
Datasheets for Datasets	Detailed reports about a given data set, including information about the specific data points and data sources. Datasheets are in-depth analyses of data sets.
Fact Sheets	Serving to improve an AI system's trustworthiness, Fact Sheets include information about a system's conformity, compliance, performance variation, security features, use of data, etc. Though information about data is included, the main focus of the Fact Sheets are the AI systems as a whole.

[46] See also for example: ai.mil/blog_09_03_21_ai_enabling_ai_with_data_cards.html
[47] See Annex IV, point 2(d).

Data Statements	A data statement characterizes a dataset, providing context to understand how experimental results generalize, how software should be deployed, and what biases may be present in software-built systems. It addresses ethical concerns such as exclusion, overgeneralization, and underexposure. Additionally, data statements enhance the developer's understanding of dataset populations, aiding in addressing scientific issues like generalizability and reproducibility[48].

Regardless of the format, data collection and processing practices must adhere to transparency and documentation requirements. Under **Article 11 of the EU's AI Act**, the provider of the AI system must deliver clear and comprehensive information to facilitate the assessment of the system's compliance. The complexity and simplicity of documentation requirements may vary depending on the size of the enterprise developing the AI system, with the European Commission specifying requirements for small and micro-enterprises.

Regarding data collection and processing, **Annex IV of the Act** stipulates that several components must be disclosed. First, information pertaining to the classification choices used in the system and the relevance of each parameter, as well as any technical trade-offs made to ensure compliance, must be provided. Data sheets must also include elements such as training methodologies and techniques, training data sets used, and descriptions of the training datasets, including their provenance, scope, and main characteristics. Additionally, information on data acquisition and selection processes, labelling procedures, and data cleaning methodologies should be included. The disclosure and clear documentation of data-related activities, including collection, governance, and processing, are essential for meeting the requirements outlined in **Article 12 of the AI Act.**

If the proposed **Fair Data Collection, Governance and Management Methodology' (FDCGM)** outlined in this document are followed, they will be taking significant steps toward meeting the EU AI Act's requirements. This methodology provides a structured approach to documenting data-related decisions and processes, though additional steps may be needed for full compliance. The next stage in ensuring **fairness-by-design** within the AEQUITAS Project involves the **Fair Model Methodology (FMM)**, which will further support organizations in meeting their documentation obligations under the AI Act.

Practical Use Case – testing of the methodology

Over the course of three weeks, ALLAI facilitated three working group sessions to trial the Fair Data Governance and Management Methodology. The exercise involved a hiring company using an AI tool designed to optimize the matching of potential employees with employers (clients). The goal was to assess the tool's dataset for fairness and ensure its compliance with the EU AI Act.

The dataset, comprising 21,000 rows collected from multiple sources between 2018 and 2023, revealed inconsistencies in data collection methods. These inconsistencies introduced potential biases, such as a demographic imbalance with only 22% female representation

[48] See: tacl_a_00041.pdf (silverchair.com) pp. 587-588.

compared to 78% male candidates. This imbalance prompted a continued debate throughout each session: should datasets reflect real-world labour market distributions, even if they mirror existing inequalities, or should they be designed to achieve more equitable representation? Ultimately, we found that historical inequities risk being perpetuated through dataset compilation if not critically examined and addressed through methodologies like the Fair Data Governance and Management Methodology. At the data feature level, we found the inclusion of features like 'dynamism', ‘communication’ and 'maturity' in candidate profiles. These features stemmed from subjective assessments creating personality evaluations and introducing potential for bias. In addition, these evaluations were marked by inconsistent scoring and missing values which compounded the risks of bias. By the end of these sessions, participants realized that decisions about data collection, feature selection, and dataset compilation are not purely technical but are value-laden and deeply rooted in societal contexts.

Several recommendations emerged from these sessions. First, there is a need for greater clarity in translating complex legal concepts from the EU AI Act into practical and implementable principles. Indeed, many participants lacked the legal expertise to interpret these mandates, underscoring the importance of making legal language more interpretable to both technical and non-technical stakeholders. The exercises also revealed a knowledge gap between technical and non-technical participants, making it difficult to sustain productive interdisciplinary dialogue. To bridge this gap, stakeholders should partake in the development of shared resources such as data dictionaries to ensure mutual understanding of key concepts. Finally, we recommend that stakeholders must be prepared to critically evaluate every data feature for potential bias and prioritize fairness over technical convenience, as required by the EU AI Act. The prevailing industry practice of trying to make datasets fit AI systems must be reconsidered. Instead, the focus should shift toward adapting both the AI system and the dataset to address societal, legal, and contextual requirements, ensuring that fairness remains central to AI development.

Template 1: Data Analysis

Part 1: Dataset Level Analysis

Overview of the Dataset				
Information	**Elements to Consider**	**Reply (use a sentence to explain technical terms in plain English)**	**Does this represent a gap / shortcoming in terms of fairness?**	**Can this gap / shortcoming be addressed to eliminate / mitigate bias / unfairness?**
Data Source (As relevant to Art 10.2.b, the contextual setting requirements under 10.4 & Annex IV.1.a)	What is the source of the data set? (the first place where the data was gathered)			
	What is the data repository? (public repository where the dataset can be found)			
	Who owns the data set? Provide the names of the data set owners.			
	Is it a primary or secondary data source?			
	Which institutes/organisations are the data set owners affiliated with?			
	Provide the contact details of the dataset owners			
	Are there any data(sharing) agreements that are or should be consolidated?			
	What are the relevant privacy and fundamental rights regulations that need to be respected when accessing the data?			
	Who has access to the data/ the			

	datasets? (internally & externally)			
Purpose of Data collection (Annex IV.1.a)	What is the purpose this data will be used for?			
	What are suitable use cases for this data set?			
	What are unsuitable use-cases for this data set?			
Data Characteristics (As relevant to Art 10.2.e & 10.3)	Volume: What is the volume of the data?			
	What volume of data is necessary for the development of the fair AI system, keeping in mind the data minimisation principle?			
	Velocity: What is the speed at which the data arrives?			
	Variety: How varied are the data types and sources?			
	What is the data sample size?			
	Were (parts of) the data set(s) unavailable?			
Data Type (10.3 & 10.4 as relevant contextual setting)	What form is this data in? e.g. structured, unstructured, semi-structured, etc			
	What are the data types present in the data set? e.g. image, audio, text, video, etc			
	Does the dataset contain any special categories of personal data under GDPR? If yes, what kind?			
Free of Errors (As relevant to Art 10.3)	Are there any inconsistencies in the data set?			
	Are there any duplicate entries?			

	Are there any missing values?			
	Are there any significant outliers or anomalies?			
	Is there version consistency?			
	Are the units of measurement consistent throughout the data set?			
	Are the data formats consistent throughout the data set?			
Data Collection (As relevant to Art 10.2.b)	How was this data collected? What collection methods were used?			
	What were the sampling methods used?			
	Were any of the sampling methods biased towards a particular group? If yes, is this bias necessary to fulfil the intended purpose?			
	What are specific and appropriate measures that address the identified bias?			
	When was this data collected?			
	What were the general criteria for including the selected data features/data points?			
	What data features/data points were excluded? Based on what criteria?			
Sufficiently Representative (Art 10.2.f, Art 3 & Art 10.4)	What demographic information and domain is the dataset aiming to represent?			
	What is the minimum sample			

	size to represent the population sufficiently and significantly?			
	Is the dataset sufficiently large to capture the variability and complexity of the underlying phenomena? Is the dataset representative of the target population or domain?			
	Are there enough samples from minority or underrepresented groups to avoid biases and ensure fairness?			
	Does the dataset reflect temporal changes, and are those changes relevant?			
	Is the dataset being updated to reflect evolving trends and changes?			
Data Processing (As relevant to Art 10.2.c, 10.2.f & 10.2.g)	Was the raw data transformed in any way? What were the data transformation methods used?			
	What data processing techniques were used? e.g. cleaning, normalisation, standardization			
	Was data labelling used? Was the labelling manual or automatic? What are the ground truths attached to each label?			
	Are the data processing techniques appropriate, given the intended			

	purposes? If no, which data processing elements are inappropriate and why?			
	What data processing tools were used?			
	Are there any inconsistencies in the data processing that might affect the completeness of the data set?			
	Have any possible biases been identified at the level of data preparation processing operations that may impact the data sets?			
	What are specific and appropriate measures that address the identified bias?			
Statistical properties (Art 10.3)	What are the measures of central tendency? Do these measures accurately represent the data set?			
	What is the variability of the data set around the central tendency? Are there any outliers or extreme values present?			
	Are there significant differences in statistical properties within the data set?			
Relevance (As relevant to Art 10.2.d)	What are the assumptions that are formulated with respect to the information the dataset is supposed to			

	represent and measure?			
	Are these assumptions appropriate, reasonable and fair in light of the AI system's intended purpose?			
	What is the rationale behind these assumptions if you look at the people that the system is going to be used for? (ANNEX IV 2(b))			
	Do the relevant formulated assumptions target a specific vulnerable group? And are those formulations likely to harm that group? If so, how can the formulations be adjusted?			
	Does the dataset cover the full scope of scenarios the AI system will encounter?			
	Will the data set provide meaningful insights or outcomes?			
	Are there any outlier cases that need to be included in the dataset?			
Data Fairness (As relevant to Art 10.2.f)	Have any possible biases that are likely to affect the health of persons been identified as a result of this examination?			
	Have any possible biases that are likely to affect the safety of persons been identified as a result of this examination?			

	Have any possible biases that are likely to negatively impact a person's fundamental rights been identified as a result of this examination? Which fundamental rights might be affected?			
	What measures can be taken to correct and eliminate the biases identified above?			

Part 2: Data Features Level Analysis

Information on Data Features				
Information	**Elements to consider**	**Reply**	**Does this present a gap / shortcoming in terms of fairness?**	**Can this gap / shortcoming be addressed to prevent /mitigate bias / unfairness? If yes, how?**
Type of datafeature (Art 10.5.b & Article 10.5.c)	Is the data feature/feature considered personal data under the GDPR?			
	Is the data feature/feature considered a special category of personal data under the GDPR?			
	Has personal data been non-reversibly anonymized?			
	Is the data variable/feature prohibited ground for discrimination (art. 21 ECFR)?			

	Could the data variable/feature infringe upon other equality, non-discrimination, fair treatment rights?			
	Could the data variable/feature be a proxy for a prohibited ground of discrimination (art. 21 ECFR), or lead to infringement of other equality, non-discrimination, and fair treatment rights?			
Scale (Art 10.2.f & 10.3)	What is the scale of the data feature?			
	Can the scale introduce bias/unfairness into the system?			
Volume (Art 10.2.f & 10.3)	What is the volume of the data feature			
	Can the volume introduce bias/unfairness into the system?			
Distribution (Art 10.2.f & 10.3)	What is the distribution of the data? Has each feature been understood in terms of its distribution?			

Assessment of the Data Features				
Information	**Elements to consider**	**Reply**	**Does this present a gap / shortcoming in terms of fairness?**	**Can this gap / shortcoming be addressed to prevent /mitigate bias / unfairness? Art. 10.2 (...) If yes, how?**

Relevance The data in the data set should be relevant to the AI's intended purpose or task that it is designed for. All elements which do not meaningfully, reasonably, purposefully and fairly contribute to the task that the system is trying to fulfil should be removed. (Art 10.2.d)	Is the information provided by the data feature necessary and proportionate for the intended purpose?			
	What is the information in the data feature supposed to measure?			
	What is the information in the data feature supposed to represent?			
	What are the assumptions that are formulated with respect to the information the data feature is supposed to represent and measure?			
	Where do these assumptions come from? Are there any historical, social, or cultural biases that underpin these assumptions?			
Free of errors (Art 10.2.f & 10.3)	Are there any duplicate entries?			
	Are there any missing elements re. the data feature?			
Complete	Does the data feature, considering its **scale**, provide a complete view/ representation of the group(s) it aims to represent?			
	Does the data feature, considering its **volume** provide a complete view/representation of the group(s) it aims to represent?			

	Is the feature a characteristic that is relevant for the geographical setting of the system			
Data Feature Interaction (Art 10.2.f & 10.3)	Are there any interactions with other features that could collectively introduce or amplify any biases or unfair outcomes that may not be apparent when considering a data feature in isolation?			
	Within the context of the data point, is the information provided by the data feature necessary and proportionate for the intended purposes? Could interactions with other features negate this?			
	Can the combination of multiple sensitive data features increase the discrimination potential?			

Part 3: Data Dictionary

Part 1 : Dataset Dictionary						
Data Source Information						
Source of the Data Set and Data Repository	**Data Owners**	**Contact details**	**Affiliated Institutions**	**Accessible to (who can access the data set?**	**Data sharing agreements**	**Privacy compliance**

Purpose of data collection						
Purpose of data collection	**Suitable use cases**	**Unsuitable use cases**	**[]**	**[]**	**[]**	**[]**

Data Set Characteristics						
Volume	**Velocity**	**Variety**	**Data Sample Size**	**List of data features**	**Unavailable parts of data set/ Missing data**	**Transformation of raw data**

Data Type						
Data form (structured, unstructured, etc)	**Data type (audio, video, text, etc)**	**Special Categories under GDPR**				

Gaps / Errors						
Inconsis-tencies	**Duplicate entries**	**Outliers and anomalies**	**Version consisten-cy**	**Measure-ment consisten-cy**	**Data format consisten-cy**	**Missing values**

Data Collection						
Data Collection Date	**Data collection methods**	**Sampling Methods**	**Sampling Bias**	**Inclusion criteria**	**Exclusion criteria**	

Sufficiently Representative						
Demogra-phic Represen-tation	**Minimum sample size for representa-tion**	**Data set sufficiently large?**	**Samples from minority groups**	**Temporal Changes**	**Evolving trends**	

Data Processing						
Techniques used	**Data Labelling**	**Appropriate ness of processing techniques**	**Data processing tools**	**Data processing aligned with stated purpose?**	**Inconsis-tenncies in data processing**	

Statistical Properties

Measures of central tendency	Variability around central tendency	Outliers	Differences in statistical properties			

Relevance						
Assumptions Represented	Assumptions: fair, reasonable & appropriate?	Rationale Behind Assumptions	Assumptions about vulnerable groups	Full scope of scenarios	Meaningful insights	Outliers considered

Data Fairness						
Data disproportionately affects certain groups?	Protections against discrimination & bias	Guiding ethical principles	Inappropriate data elements	Steps to rectify inappropriate data elements	Fairness of data set	Steps to rectify unfairness of data set

Part 2: Data Features Dictionary

	Information per data feature (collected in part A)					
	Data type (Personal data must be non-reversibly anonymized)	**Scale (value?)**	**Volume**	**Statistical property**	**Geographic elements**	**Proxy for prohibited grounds for discrimiation**

[NAME DATA FEAUTURE]						
	Assessment of datafeature (collected in part B)					
	Relevance for intended purpose (include justification)	**Errors / gaps**	**Complete**	**Reasonable and fair assumptions**		
	[…]	**[…]**	**[…]**	**…**	**…**	**…**

Data Features Evaluation				
Fairness gap /shortcoming identified (dataset level)	**Unfairness/Bias prevention measure**	**Unfairness/Bias mitigation measure**	**Describe 'level' of mitigation**	**Solution**