CEN–CENELEC JTC21 and the EU AI Act Article 10 (Data)

James Davenport
Convenor JTC 21 WG 3

University of Bath

27 June 2025

EU AI Act Article 10

10.1 High-risk AI systems which make use of techniques involving the training of AI models with data shall be developed on the basis of training, validation and testing data sets that meet the quality criteria referred to in paragraphs 2 to 5 . . .

10.2 Training, validation and testing data sets shall be subject to data governance and management practices appropriate for the intended purpose of the high-risk AI system. [. . .]

10.3 Training, validation and testing data sets shall be relevant, sufficiently representative, and to the best extent possible, free of errors and complete in view of the intended purpose. [. . .]

10.4 Data sets shall take into account, to the extent required by the intended purpose, the characteristics or elements that are particular to the specific geographical, contextual, behavioural or functional setting within which the high-risk AI system is intended to be used.

10.6 For the development of high-risk AI systems not using [training] paragraphs 2 to 5 apply only to the testing data sets.

10.2 (I)

Training, validation and testing data sets shall be subject to data governance and management practices appropriate for the intended purpose of the high-risk AI system. Those practices shall concern in particular:

(a) the relevant design choices;

(b) data collection processes and the origin of data, and in the case of personal data, the original purpose of the data collection;

(c) relevant data-preparation processing operations, such as annotation, labelling, cleaning, updating, enrichment and aggregation;

(d) the formulation of assumptions, in particular with respect to the information that the data are supposed to measure and represent;

10.2 (II)

(e) an assessment of the availability, quantity and suitability of the data sets that are needed;

(f) 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;

(g) appropriate measures to detect, prevent and mitigate possible biases identified according to point (f);

(h) the identification of relevant data gaps or shortcomings that prevent compliance with this Regulation, and how those gaps and shortcomings can be addressed.

'Training, validation and testing data sets shall be relevant, sufficiently representative, and to the best extent possible, free of errors and complete in view of the intended purpose. They shall have the appropriate statistical properties, including, where applicable, as regards the persons or groups of persons in relation to whom the high-risk AI system is intended to be used. Those characteristics of the data sets may be met at the level of individual data sets or at the level of a combination thereof.

10.5 (not being standardised currently): I

To the extent that it is strictly necessary for the purpose of ensuring bias detection and correction in relation to the high-risk AI systems in accordance with paragraph (2), points (f) and (g) of this Article, the providers of such systems may exceptionally process special categories of personal data, subject to appropriate safeguards for the fundamental rights and freedoms of natural persons. In addition to the provisions set out in Regulations (EU) 2016/679 and (EU) 2018/1725 and Directive (EU) 2016/680, all the following conditions must be met in order for such processing to occur:

(a) the bias detection and correction cannot be effectively fulfilled by processing other data, including synthetic or anonymised data;

(b) the special categories of personal data are subject to technical limitations on the re-use of the personal data, and state-of-the-art security and privacy-preserving measures, including pseudonymisation;

10.5 (not being standardised currently): II

(c) the special categories of personal data are subject to measures to ensure that the personal data processed are secured, protected, subject to suitable safeguards, including strict controls and documentation of the access, to avoid misuse and ensure that only authorised persons have access to those personal data with appropriate confidentiality obligations;

(d) the special categories of personal data are not to be transmitted, transferred or otherwise accessed by other parties;

(e) the special categories of personal data are deleted once the bias has been corrected or the personal data has reached the end of its retention period, whichever comes first;

(f) the records of processing activities pursuant to Regulations (EU) 2016/679 and (EU) 2018/1725 and Directive (EU) 2016/680 include the reasons why the processing of special categories of personal data was strictly necessary to detect and correct biases, and why that objective could not be achieved by processing other data.

10.5: why not standardised?

1. It's a conditional permission, not a requirement
2. Lack of prior art in general, and specific knowledge in the committee

(In particular, how do you standardise "cannot be effectively fulfilled " and "strictly necessary")

But help would be welcomed.

Data Standard (version 20250625)

§6 Dataset governance (AI Act 10.2) (prior planning etc.; interaction with QMS)

§7 Methodology applicable to raw data

§7.10 Poisoned data (crossreference Cybersecurity)

§7.11 Debiasing Data (crossreference Bias)

§8 Methodology applicable to data annotations

§8.11 Poisoned annotations (crossreference Cybersecurity)

§8.12 Debiasing Annotations (crossreference Bias)

§9 Dataset release (terminology from ISO 12792 Transparency)

§10 Dataset maintenance (including retirement)

§11 Use of datasets in AI; especially semantic independence of testing data

Bias Standard (version 20250625)

§6 Manifestation of bias in AI systems and datasets

§7 Examples of bias

§8 Bias management requirements (interactions with Risk and QMS)

§9 Assessment of bias

§9.2 Process for assessing bias in datasets

§9.4 Measures of bias (help welcomed)

§10 Specifications for methods to reduce bias

§10.2 Reduction of bias in raw data

§10.3 Reduction of bias in data annotations

§11 Specifications for methods to reduce the impact of bias

§11.2 Impact reduction of bias from raw data

§11.3 Impact reduction of bias from data annotations

Issues

- Need more input, especially on data set governance and life cycle: ask your national standards body to nominate you to JTC21/WG3.

Or send me suggestions/ papers `masjhd@bath.ac.uk`

- EU Data Act (standards being developed in JTC 25)
- Cybersecurity. The AI Act applies even where Cyber Resilience Act doesn't, so we can't easily rely on JTC 13.

Bibliography I