EDPB public consultation: Guidelines 02/2026 on Anonymisation

Submission date: 11th of August 2026

Contributor: Michael Kühne-Schlinkert (Katulu GmbH, Germany - Provider for AI privacy and security solutions and technical AI assessments)

Cover statement

Declared interest, stated first. Katulu GmbH performs technical assessments of whether personal data can be recovered from trained machine-learning models. We would benefit commercially if these Guidelines required such assessments, and more so if they required independent ones. They should not, and the comments below do not ask for either. Point 4 in particular asks for materially less than would serve our interest.

The four comments concern paragraphs 35 and 41 and the treatment of trained models. Each is drafted to be useful to a controller performing a self-assessment, because a comment that only helps the commenter is worth nothing to the Board.

Comment #1 on ¶41: specify what "documentation including testing" contains

Observation: ¶41 expects documentation covering the tests performed on the allegedly anonymous data, and its retention. It does not specify what that documentation must contain. Documentation with no defined minimum content cannot be assessed - not by a supervisory authority, and not by a counterparty to whom it is handed.

Suggestion: The Guidelines should state a minimum content for such documentation:

- the threat model, including the assumed capabilities and access level of the adversary
- the exact version of the dataset or model tested
- the methods applied
- the date of testing
- the results, with the metrics used
- the stated limitations of what was tested

Reasoning: Without these, a test result is neither reproducible nor comparable between controllers. With these minimum requirements, an assertion becomes verifiable evidence. Comparability is what makes the Article 5(2) accountability obligation operable in practice, and it is achieved at no additional cost to a controller who has already done the testing.

Comment #2 on ¶35: name triggers rather than an interval

Observation: ¶35 recommends periodic re-assessment as good practice, without indicating periodicity. In practice this produces either an arbitrary annual cycle or nothing at all.

Suggestion: The Guidelines should identify event-based triggers instead of an interval:

- a material change to the dataset or model
- a change in the conditions of access or publication
- a material advance in the state of the art of the relevant attacks.

Reasoning: Re-identification risk does not simply increase over time. It increases when a new attack is published or when access conditions change, which is what ¶35 itself describes when it notes that likelihood "typically increases over time". A trigger tracks the actual risk. An interval tracks the financial year. Triggers are also less burdensome for controllers whose datasets and models are stable.

Comment #3: trained models are not datasets, and the Guidelines should say how they differ

Observation: The Guidelines are drafted for datasets. The three criteria (singling out, linkability, inference) apply to trained models in principle, but assessing them is a different exercise. A dataset is examined by reading it. A model is examined by attacking it.

The Board has already recognised the difference. EDPB-EDPS Joint Opinion 2/2026, adopted 10 February 2026, records at ¶42 that *"it might for instance prove technically difficult to remove personal data that is retained in the AI system or model."* Personal data retained in a model is not reached by an assessment written for records in a table.

Suggestion: A dedicated section or annex should set out how the three criteria are to be assessed for trained models, and name the relevant classes of attack: membership inference, verbatim extraction of training content, model inversion, reconstruction attacks, and attribute inference.

Reasoning: Singling out and linkability each have named attacks in the list above. Inference (third criteria from the Guideline) currently has none. Attribute inference is the attack that corresponds to it, whether a sensitive attribute of an individual can be inferred more accurately with access to the model than from population data alone.

One qualification belongs in the same context, because a Guideline that named the attack without it would invite results that overstate risk. A model predicting that people with certain characteristics tend to have a certain condition has learned a population regularity, which is what a model is for. Therefore, the measure should be the increase in inference accuracy attributable to model access, measured against a population baseline computed without the model (not raw inference accuracy).

The same vocabulary already appears in material the EDPB has published. The Support Pool of Experts report *AI Privacy Risks & Mitigations - Large Language Models* recommends *"structured testing against state of the art attacks such as attribute and membership inference, exfiltration, regurgitation of training data, model inversion, or reconstruction attacks"*, and lists among the evidence to retain *"evidence of the model's theoretical resistance to re-identification techniques, including metrics, testing reports, and analysis of attack resistance."* Using the same vocabulary in the Guidelines as in this SPE report would spare controllers the task of reconciling two differently worded expectations.

Comment #4: independence as a factor in evidential weight, not as a requirement

Observation: The Guidelines are silent on who performs the assessment. That is correct: the GDPR nowhere requires an independent assessment, and the Guidelines should not create one.

Suggestion: The Guidelines should state only that the independence of the assessor from the development of the dataset or model is a factor in weighing the evidential value of an assessment - in particular where the assessment is presented to a third party who cannot repeat it.

Reasoning: The same distinction is drawn by the Swiss financial supervisor FINMA, which in its supervisory notice 08/2024 expects a clear separation between the development of AI applications and their independent review, while requiring no external provider. Independence of the reviewer from the developer is achievable inside an organisation and is a recognised control in every other assurance discipline.

A European standards body has specified the same distinction. ETSI TS 104 008 (January 2026), on continuous auditing-based conformity assessment, holds at Clause 5.3.4 that independence and objectivity - *"whether internal or external"* - are *"not only an assumption, but a fundamental requirement"*, fostered through clearly defined roles, *"strict reliance on objective evidence derived from metrics and measurements"*, and *"the option of third-party assessment modes."* Independence is treated there as a property of the arrangement, not as a requirement to engage an external provider, which is exactly the distinction proposed here.

This proposal creates no obligation and no market for assessors. It records only what is already true: a self-assessment produced by the same team that built the model carries different evidential weight from an independent one. As stated in the cover, we would have benefited from asking for more, and have deliberately not done so.

Summary

#	Paragraph	Asks for	Burden on controllers
1	¶41	A minimum content for test documentation	**None** - it structures work already done
2	¶35	Event triggers instead of an interval	**Reduced** for stable datasets
3	New	Model-specific guidance naming five attack classes, so that each of the three criteria has one	**Reduced** - removes ambiguity
4	New	Independence as one factor in evidential weight	**None** - no obligation created