Spolek pro ochranu osobních údajů

Comments on the EDPB draft of Gudelines 03/2026

We welcome the opportunity to provide feedback on the recently published EDPB draft of Guidelines 03/2026 on web scraping in the context of generative AI.

We appreciate the considerable work undertaken and the effort to harmonise the training of AI across EU. This is particularly necessary given the EU's lag in this area. We need therefore to seek pragmatic solutions. Even though we in the EU have opted for a 'regulation first' approach, we should not give up on at least trying to ensure an acceptable environment for the development of large European LLM models (given that we are, after all, more successful in developing at least smaller, specialised models).

General comments:

1. Unfortunately, we have some doubts as to whether the draft guidelines can unequivocally be described as "future-proof". By focusing on current forms of training, which may soon become outdated, the guidelines themselves may become obsolete within a very short time.

   In particular, we note the absence of the EDPB's views on several scenarios that already exist or are likely to arise in near future, such as:

   a. Recursive self-improvement of models

   b. Retrieval Augmented Generation and enriching models with data obtained in the course of their normal use

   c. The emergence of AGI – Artificial General Intelligence – which is likely to blur the distinction between an AI model and an AI system.

   The guidelines follow the logic of the GDPR, which is, of course, the correct approach. However, that approach may fall short where the EU aims to develop competitive LLMs of its own. In that context, formal application of the GDPR may become a real obstacle. The GDPR's requirements should therefore be interpreted, for the purposes of AI training, as incorporating a "reasonable efforts" criterion, in a manner consistent with the GDPR's risk-based approach and the broader principles reflected in the AI Act.

2. This is particularly important in cases involving data that identify data subjects only indirectly, or where identification is possible but would require active and deliberate steps, for example by combining multiple datasets.

Spolek pro ochranu osobních údajů / Data Protection Association
with its registered seat at the address Hellichova 458/1, Malá Strana, 118 00 Praha 1 / Prague 1,
registration with public register: Municipal Court in Prague, file No. L 61299
Business ID No.: 03493679, e-mail: spolek@ochranaudaju.cz, tel.: +420 730 677 644

3. It should also be recognised that, in some cases, filtering only at the output stage (combined with reasonable efforts to filter the relevant data out of the input datasets) is the only practicable method of ensuring compliance with the GDPR.

4. It would also be appropriate to give greater consideration to the interplay with other areas of law that also address issues concerning the lawfulness of input data, such as copyright law. In particular, this concerns the relationship with the text and data mining (TDM) exceptions in Articles 3 and 4 of Directive (EU) 2019/790 (see below) and the requirements of Article 53 of the AI Act.

5. The emphasis on synthetic data and their use may be inappropriate in some cases, particularly where the very purpose of the system or model is to process personal data, or where even minor deviations caused by converting data into synthetic data may cause serious problems (for example, in models used in healthcare, where retrospectively identifying faulty model behaviour may be very difficult without using "real" data).

Comments on individual paragraphs:

Paragraphs 27 and 36: Although we agree with the EDPB's conclusions, we note that, for LLMs as such, it may be difficult to identify the "final" purpose of processing within the system in which they will be deployed. For these models, the purpose of processing may in fact be the very broad purpose of "developing a suitable LLM for deployment in common processing use cases and a wide range of AI systems". It would therefore be useful to clarify further how to proceed where the purpose is formulated in such general terms.

As regards **paragraph 28**, we welcome the EDPB's position that the balancing test should be carried out in relation to the dataset as a whole, rather than individual items of data within it.

Paragraphs 30 and 31: We would welcome the EDPB's views on the relationship between these obligations and those under Article 53 of the AI Act.

Paragraphs 34 and 38: Although we agree that collecting data spanning a period of 20 years may be problematic, we must point out that, for some systems, the collection and processing of personal data may be their very purpose, or that processing such data may not be problematic even where they concern past events. Examples include queries such as "Who was the President of the Czech Republic in 2000?" This aspect would benefit from a more detailed and flexible approach, rather than being treated as a blanket prohibition.

Paragraph 37: Use of synthetic data. As noted above, we wish to point out that, while the use of synthetic data may be appropriate in some cases, it may pose risks in many others. This applies particularly to their use in healthcare, where even the slightest deviation or inaccuracy may pose a significant risk. Given the "black box" nature of AI systems, such a deviation may then no longer be easy to eliminate. From this perspective, the use of synthetic data may also conflict with healthcare legislation and product requirements applicable to data processing systems in healthcare.

Spolek pro ochranu osobních údajů

Paragraph 37: As regards the requirement to "*exclude from the collection websites which clearly oppose the scraping*", it would be necessary to align these requirements with the opt-out rules under Articles 3 and 4 of the DSM Directive (Directive (EU) 2019/790).

Paragraphs 39 and 42: We agree that data accuracy must be ensured and that prioritising data from trustworthy sources has numerous advantages. However, the text somewhat overlooks the fact that data may be accurate for the purpose for which they are processed even if they are objectively inaccurate (for example, data collected for systems designed to research disinformation, which require the source data in their original, unaltered form).

Paragraph 66: Although we generally agree with the views expressed here, we must also point out that excluding certain categories of data or certain groups of individuals may lead to unintended bias when training AI models.

The opt-out lists (or other forms of model unlearning) referred to in point (f) also present a much more complex issue in practice, particularly in light of the definition of personal data. In many cases, a particular data subject can be identified only through context, by bringing together a range of data about them which, only when taken as a whole, establish which data subject is actually concerned. Even if, for example, an individual data subject objects, it may not be clear which data should be excluded. One example is an objection raised by a person named Jan Novák (a very common Czech first name and surname): it may not always be clear which Jan Novák, among the many people with that name whose data are processed in the system, the objection concerns.

Example 6: It is not entirely clear what this example is intended to illustrate. It would be useful to explain it in more detail, particularly which specific measures the EDPB has in mind and in what context.

Paragraph 67 (and, to some extent, paragraph 74(ii)): Although we welcome the EDPB's position on limiting the scope to "intended scraping", which probably builds on the proposed amendment to Article 9 of the GDPR in the Data Omnibus, we must point out that this position is incompatible with the case-law of the CJEU. Although this presents a serious obstacle to training AI models in the EU, in our view it can only be resolved through an amendment to the GDPR or a change in the CJEU's (very strict) case-law. The EDPB could, however, play a very positive role in this regulatory dialogue with the Commission and other stakeholders.

Spolek pro ochranu osobních údajů

We are grateful for the opportunity to provide our comments on the draft Guidelines. We would like to thank the EDPB very much for its efforts to clarify various issues related to the processing of personal data. We hope that our comments will help in formulating the final version of the guidelines.

Prague, 29.10.2026

JUDr. Vladan Rámiš, Ph.D.
Chairman of the Committee
Data Protection Association

Mgr. František Nonnemann
Vice-Chairman of the Committee
Data Protection Association