Contribution to the public consultation on EDPB Guidelines 03/2026 on web scraping in the context of generative AI

Responsibility Culture and the effective implementation of data protection safeguards

Ines Caloisi

Introduction

I welcome the opportunity to comment on Guidelines 03/2026. This contribution draws on a sociological perspective on artificial intelligence, social responsibility and Responsibility Culture, understood here as the organisational capacity to translate responsibility into shared practices, documented decisions and continuous evaluation. In legal terms, this perspective aims to support the practical implementation of the accountability principle (Article 5(2) GDPR), the responsibility of the controller (Article 24 GDPR) and data protection by design (Article 25 GDPR).

The Guidelines provide important clarification on the responsibilities associated with collecting personal data from publicly accessible sources for generative AI development. This contribution proposes three practical clarifications intended to support their effective implementation: documenting organisational responsibility, assessing data subjects' reasonable expectations, and evaluating the effectiveness of safeguards. The proposals are intended to clarify existing obligations, not to introduce new ones.

1. Make organisational responsibility traceable

Relevant paragraphs: 14, 16–21 and 32

Paragraph 14 acknowledges that it is "not obvious" how to meet the accountability principle when organisations carry out scraping or obtain scraped datasets whose personal data content may not be fully known, while recalling that compliance with Article 5(2) GDPR must nonetheless be met by the controller. Paragraphs 16–20 clarify the respective roles of the organisations involved, and paragraph 21 states that the controller must consider the datasets it intends to use and the types of websites it would like to scrape.

Further practical guidance could help organisations connect these legal roles with identifiable decision-making responsibilities. In complex supply chains, responsibility may become fragmented between those selecting sources, collecting data, obtaining datasets from third parties (paragraphs 19 and 32) and authorising their use for training.

The Guidelines could recommend that controllers document:

the functions responsible for approving collection criteria and source selection;

the information examined before obtaining or using a scraped dataset, including its provenance and the conditions under which it was collected;

the reasons for accepting, restricting or rejecting particular sources;

procedures for escalating uncertainty and reviewing decisions when new risks emerge.

These arrangements should support, not replace, the factual assessment of controller and processor roles described in paragraph 16. They should be proportionate to the size of the organisation and to the scale and risks of the processing, and can be integrated into existing documentation such as the record of processing activities and the data protection impact assessment referred to in paragraph 34.

Suggested addition (after paragraph 21)

“To support accountability under Articles 5(2) and 24 GDPR, controllers should document how decisions concerning source selection, collection criteria and the use of scraped datasets, including datasets obtained from third parties, are authorised and reviewed within their organisation. Such documentation should identify responsible functions, the information assessed and the procedures for addressing uncertainties or newly identified risks, in a manner proportionate to the scale and risks of the processing.”

2. Assess reasonable expectations in their social context

Relevant paragraphs: 45, 63–64 (including Examples 4 and 5) and 66(c)

The Guidelines already contain important elements on this point. Paragraph 63 states that the mere fulfilment of information obligations is not sufficient in itself to establish that data subjects can reasonably expect a specific processing. Paragraph 64 recognises that people cannot be considered to expect processing “in all situations, for all purposes”, and lists the characteristics of the data subjects among the relevant criteria. Paragraph 45 clarifies that the absence of a robots.txt file does not amount to consent, and paragraph 66(c) acknowledges that data may be freely accessible without data subjects being reasonably aware of that accessibility.

This contribution suggests applying these principles consistently in the examples. Awareness that third parties may access published content does not necessarily imply an informed expectation that personal data will be collected at scale, combined with other sources and used to train generative AI models.

Example 4 currently concludes that data subjects “can reasonably expect” third parties to scrape their data to develop AI models, on the basis that the platform is freely accessible, contains no prohibitions and communicates the possibility of scraping. Example 4 would benefit from further qualification to reflect the contextual assessment described in paragraph 63: a platform’s communication may contribute to the assessment, but its practical significance depends on whether it is prominent, understandable and appropriate to the users concerned. The context in which information was shared, differences in digital literacy and the presence of vulnerable persons are also relevant. The absence of technical restrictions should not, by itself, establish that data subjects reasonably expect AI training.

Suggested revision of Example 4

“A person uploads their data to a content sharing platform that is freely accessible and does not contain any prohibitions against web scraping, and the platform communicates the possibility of the content being scraped to its users. Such communication may contribute to the reasonable expectations of data subjects that third parties may scrape that data to develop AI models, provided that it is prominent, clear and appropriate to the users concerned, taking into account in particular the presence of minors or other vulnerable persons among them.”

Suggested addition (paragraph 64)

“When assessing reasonable expectations, controllers should distinguish between a data subject’s awareness of public accessibility and their reasonable expectation of subsequent processing for generative AI development. Relevant considerations include the context of disclosure, the prominence and comprehensibility of the information provided, and the characteristics of the data subjects concerned, including their vulnerability. Neither public accessibility nor the absence of technical restrictions should, on its own, be treated as sufficient evidence of such expectations.”

3. Evaluate safeguards through their practical effects

Relevant paragraphs: 33, 60, 65–66 and 76

Paragraph 76 expressly recommends that measures adopted to prevent the collection and dissemination of special categories of personal data be accompanied by regular verification and assessment of their effectiveness, and adjusted where they prove ineffective. In this section, regular verification is expressly addressed in the context of Article 9 GDPR. Paragraphs 65–66, which describe the mitigating measures relevant to the balancing test under Article 6(1)(f) GDPR, could usefully make this expectation equally explicit, without implying that verification duties are otherwise absent.

Yet these measures are decisive: according to paragraph 65, if sufficient mitigating measures are not taken, the processing cannot be based on legitimate interest. It would therefore be consistent to clarify that their effectiveness should also be verified, so that the existence of a policy or procedure is not treated as sufficient evidence that it protects data subjects in practice.

For transparency and the exercise of rights (paragraphs 33 and 66(e)–(f)), evaluation could consider whether people can find and understand the relevant information, identify the organisation responsible, and use an accessible channel to submit requests. This is particularly relevant given that paragraph 60 recognises the practical difficulty for data subjects in exercising their rights when data are scraped by third parties. For measures protecting persons at particular risk (paragraph 66(a)) and vulnerable persons (paragraph 37), organisations could examine recurring failures, difficulties reported by affected people and whether corrective action prevents repetition.

Possible indicators include:

accessibility and clarity of information and request channels;

response times and the effectiveness of action taken;

recurring complaints or failures affecting particular groups;

documented changes following testing, requests or incidents.

These indicators should be proportionate to the processing and its risks.

Suggested addition (after paragraph 66)

“As for the measures referred to in paragraph 76, the mitigating measures adopted in the context of the balancing test should be subject to regular verification of their practical effectiveness for the data subjects concerned, including the accessibility of information and rights-exercise mechanisms, the effectiveness of corrective action and recurring difficulties affecting vulnerable persons. Where measures prove ineffective, controllers should implement additional or different measures and document the changes made.”

Closing remarks

A Responsibility Culture perspective can help organisations embed data protection in routine decisions throughout data collection and AI development, giving concrete form to the accountability principle.

The proposed clarifications build on principles the Guidelines already contain: the controller’s duty to consider the datasets it uses (paragraph 21), the insufficiency of information alone to establish reasonable expectations (paragraph 63), and the regular verification of the effectiveness of measures (paragraph 76). Applying them consistently would strengthen the connection between accountability, organisational practice and the effective protection of data subjects, and help make responsibility demonstrable through decisions, accessible procedures and evidence of improvement.