Guidelines 03/2026 on web scraping in the context of generative AI

Version 1.0

Adopted on 07 July 2026

Version history

Version	Date	Adoption information
Version 1.0	07 July 2026	adoption of the guidelines for public consultation

Executive summary

Web scraping is a commonly used technique for extracting large amounts of data from publicly available web services. It is a large-scale processing activity and often occurs without data subjects being aware of it. This poses particular risks to individual's rights and freedoms, and in particular their right to protection of personal data.

These guidelines focus on situations where data is scraped from internet sources, in the context of training generative AI. They cover situations were an organisation either scrapes data from internet sources external to the organisation themselves, or by contracting another party. They are limited to web scraping done by private entities.

The GDPR applies to web scraping when it includes personal data processing operations, such as collection, storage, organisation and retrieval. Compliance with certain GDPR requirements can be challenging when web scraping collects data that include personal data.

Different organisations may be involved in the data collection process for the creation of an AI training dataset, with varying degrees of involvement. The qualification of the organisations involved in each processing as controllers, joint controllers or processors should be analysed on a case-by-case basis.

The controller must comply with the purpose limitation principle, as well as the transparency principle. For the latter, this includes the obligation to inform the data subject about the processing of their personal data. Depending on how the data processing is precisely designed, the controller might not have to inform data subjects individually about the web scraping if the provision of the information proves impossible or would require disproportionate effort.

To comply with the principle of data minimisation**,** the controller should implement measures to ensure that only information necessary for the intended purpose is collected. Such measures could be: before the collection of data, considering using synthetic data instead of personal data; defining precise collection criteria; undertaking a data mapping and inventory process; apply filters to exclude the collection of certain categories of data; exclude from the collection websites that structurally contain certain types of data and websites which clearly oppose the scraping. After the collection of data, the controller could apply syntax-based filtering mechanisms and, where feasible, replace some or all real data with synthetic data, or anonymise or pseudonymise the data.

To ensure the accuracy of the personal data, it is recommended that the controller scrape from reliable sources, timestamp the data and validate the data before using them in AI training.

The legal basis of legitimate interest (Article 6(1)(f) GDPR) is often used by private entities for scraping for generative AI training purposes. In order for the application of Article 6(1)(f) to be lawful, three cumulative conditions must be fulfilled: the pursuit of a legitimate interest by the controller or by a third party; the necessity to process personal data for the purposes of the legitimate interests pursued; and that the interests or fundamental rights of the person concerned by the data protection do not take precedence over the legitimate interest of the controller or a third party (the so-called "balancing test"). As a part of the balancing test, controllers can implement technical or organisational measures to mitigate any identified risks in the context of web scraping for generative AI.

Such mitigating measures could include, among others: ensuring that certain data categories are not collected or that certain sources are excluded by default from data collection; limiting

the collection to freely accessible data; establishing safeguards that contribute to increased transparency; facilitating the exercise of individuals' rights; deleting or anonymising personal data as soon as possible and pseudonymise personal data. Other additional measures relating to the subsequent stages of the AI development phase, such as including measures to limit the risks of memorisation, regurgitation or attack of AI models or systems are also relevant.

Processing of special categories of data is in principle prohibited. For the web scraping entailing the processing of special categories of personal data, a derogation in Article 9(2) GDPR is required, in addition to a lawful basis in Article 6 GDPR. The EDPB considers that the Court's ruling GC & Others (C-136/17) can be relevant for the incidental and residual collection by a controller of special categories of personal data in the context of web scraping for the training of an AI model, '*within the framework of his responsibilities, powers and capabilities,*' if the controller implements technical and organisational measures to prevent the collection and the dissemination of data. The EDPB recalls that the controller should carry out a case-by case analysis of whether the reasoning of the Court's ruling can be applied to a specific processing activity.

Table of Contents

The European Data Protection Board

Having regard to [Article 70 (1)(e) of the Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, (hereinafter "GDPR"),

Having regard to the EEA Agreement and in particular to Annex XI and Protocol 37 thereof, as amended by the Decision of the EEA joint Committee No 154/2018 of 6 July 2018,

Having regard to Article 12 and Article 22 of its Rules of Procedure,

Has adopted the following Guidelines:

1 Introduction and scope

1 These guidelines aim to clarify the legal and technical implications of web scraping for generative artificial intelligence (AI) development under the GDPR. Generative AI is a technology aiming to create new content by learning patterns from existing data. It uses specialised machine learning models designed to produce a wide and general variety of outputs such as text, image or audio.

2 Generative AI models are built through different phases, including training and, where relevant, fine-tuning. Training generative AI models requires vast amounts of data from diverse sources. In the context of AI training, web scraping is used to collect large datasets from publicly accessible sources on the internet, which may be employed not only to train new generative AI models but also to fine-tune existing models for specific use cases.

3 The following are typical situations regarding data scraping for the purpose of training AI models.

- An organisation scrapes data from the internet themselves or by contracting another party, in the context of the development of generative AI.
- An organisation obtains and reuses a dataset that has already been scraped by another organisation (data broker).

4 These guidelines will focus on these scenarios. The guidelines will not address the role of data brokers who scrape data and just hold data sets available for other organisations, and whose intention is not to train AI models themselves.

5 Furthermore, these guidelines focus on situations in which personal data is acquired (scraped) from sources external to the organisation and do not address processing of an organisation's own personal data. This does not mean that the excluded situations are not subject to data protection legislations.

6 As web scraping for the development of generative AI is mainly performed by private sector bodies, these guidelines will also focus only on web scraping done by private entities.

2 Definition of web scraping

7 Web scraping is a commonly used technique that uses automated tools for extracting and storing information from publicly available web services, for example public registers, open-data portals, news outlets, social media, discussions' forums, and personal websites ('blogs'). Many of these websites are likely to contain personal data.

Types of web scraping and content

8 An overarching distinction can be made between targeted scraping and untargeted scraping, depending on the specificity and restrictiveness, of the collection (and not the technical tools used):

> **Targeted scraping** is performed according to collection criteria that can be more or less restrictive. The criteria could be, for example: all URLs ending in europa.eu, URLs with content written in a specific language or covering a specific topic.
>
> **Untargeted scraping** (such as web crawling without collection criteria) is done on more unrestricted basis, where the software is allowed to explore and follow each discovered link and continue with its exploration. This can be performed by using a web crawler, which automatically updates the list of URLs to be visited and store the server replies. This is done with 'spiders', small software programs that receive their instructions (tasks) in advance. These tasks could be: "add all the URLs you encounter while crawling the listed URLs". The task could also be limiting the list or URLs to visit, for example "follow all links as long as you stay within the domain www.edpb.europa.eu". Depending on the instructions to the spiders, a relatively short list of URLs can expand significantly over time as the crawler continuously discovers and adds new URLs to its queue of pages to be visited. This may result in the exploration of a large portion of the internet to be searched as a result of the constant addition of relevant URLs. This increases the risk that the controller has limited knowledge of what personal data are collected and processed.

9 An additional distinction can be made between static and dynamic content in websites. It is common practice to combine both static and dynamic content on the same site or webpage.

> **Static content** does not change due to users' actions (e.g. the text of a given post in a blog). When scraping static content from web pages, the target data resides directly in the HTML of a webpage, meaning it can be extracted simply by fetching and parsing the page. Most scrapers targeting static websites rely on the direct structure of HTML to locate the needed data.
>
> **Dynamic content** can change to respond to users' actions (e.g. loading more content when a user scrolls down a page). When scraping dynamic content from webpages, web crawlers need to render the page internally, mimicking a human users' interaction with the page. These scrapers wait for the page to render the additional content and then read and extract the dynamically rendered data.

10 Websites change their structure frequently. Scrapers might need regular monitoring and updates to continue functioning correctly (e.g., when updated data is needed for the model to perform accurately).

The different steps in the process of web scraping

11 The process of web scraping can typically be sequenced into several separate steps.

> The first step relates to **defining data collection criteria**. It involves defining the relevant sources which will be scraped depending e.g. on the purpose of

the following AI-development, as well as other criteria to exclude the collection of unnecessary personal data for the processing. If instructed to do so in line with the selection criteria, the scraper might update its list of URLs to be visited with links found at the source. This step also implies implementing measures to exclude certain sources from the collection (e.g. websites which oppose web scraping with measures such as robots.txt or CAPTCHA). For more details, see part 3.2.2 on data minimisation.

The second step relates to **"extraction**", during which data from the selected sources is transferred to the designated storage location.

The third step is the "**cleaning**", as the raw data transferred by the automated process will often need to be cleaned to improve data quality (e.g. avoiding duplicated data, removing irrelevant HTML tags or code and normalizing formats (e.g. dates, numbers), delete unnecessary data).

The fourth step is the "**structuring and storing**", during which the organisation building the training dataset might apply some further rules to refine the data by excluding low quality Internet content. After this refinement step, the remaining data will be structured into a usable format.

3 Web scraping and the GDPR

3.1 General observations

12 The GDPR will apply to web scraping when it includes processing operations, such as extraction, cleaning, structuring and storing of personal data, e.g., data that directly identifies an individual (e.g. names, pictures or contact details) or data that indirectly identifies an individual (i.e. information that can be linked to an identifiable individual considering all the means reasonably likely to be used).

13 Compliance with certain GDPR requirements can be challenging when web scraping is used to collect personal data, such as identifying a lawful basis for processing under Article 6 GDPR, meeting the obligations for processing of special categories of personal data under Article 9 GDPR, complying with the obligation to ensure transparency towards data subjects (under Article 5(1)(a), 12-14 GDPR), and complying with, but not limited to, principles such as purpose limitation (Article 5(1)(b) GDPR), data minimisation (Article 5(1)(c) GDPR), fairness (Article 5(1)(a) GDPR), and accuracy (Article 5(1)(d) GDPR). The EDPB recalls that in case of 'mixed' datasets (including both personal and non-personal data), GDPR applies to the processing of personal data. Organisations must carefully consider these requirements. They should conduct their scraping activities in a way that enables them to ensure compliance with these principles and requirements, and include safeguards that protect personal data to facilitate compliance with the GDPR and avoid infringing individuals' rights.

14 Collecting vast amounts of personal and potentially sensitive data[1], which usually happens without the data subjects' knowledge, could provide the controller with very detailed information about data subjects, irrespective of whether it is intentional or not. Considering

[1] On "sensitive" data see Article 29 Working Party Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679, wp248rev.01, endorsed by the EDPB, p. 9-10.

the technical aspects and goals of web scraping, it might be challenging for the organisation carrying out the scraping, and an organisation obtaining the scraped data, to determine exactly what personal data, and related to which data subjects, are included in the data set. It is thus not obvious how to meet the principle of accountability laid down in Article 5(2) GDPR when processing such data sets for AI development. Nonetheless, compliance with regard to Article 5(2) GDPR must be met by the controller. Furthermore, if an AI model memorizes and reproduces data included in a training data set, it poses additional risks to the privacy of individuals and the data subjects' rights. The information memorised by the AI model could be used directly for arbitrary purposes. Other risks include the fact that AI models could infer additional information about individuals by using the training data about the data subjects as well as additional information about other data subjects. The reasoning capability of the model based on the training data can also pose an additional risk to individuals.

15 In summary, web scraping can be a powerful tool for generative AI development, but it poses significant data protection risks. These guidelines are intended to assist controllers in navigating the requirements of the GDPR when using web scraping for generative AI training.

3.2 Guidance on specific provisions in the GDPR

3.2.1 Controller/joint controllers/processor

16 Different organisations may be involved in the data collection process for the creation of an AI training dataset, with varying degrees of involvement. The qualification of the organisations involved in each processing (i.e. controllers and processors), within the meaning of the GDPR, should be analysed on a case-by-case basis. It should stem from an analysis of the factual elements or circumstances of the case, including the way in which the data processing at hand is organized.[2] However, some general considerations can be provided:

17 The organisation performing the scraping is not necessarily the controller under the GDPR.

18 There might be cases where an organisation (AI developer) that is planning to train an AI model contracts another party to perform the scraping (scraper). If the scraper creates a training data set on behalf of the AI developer, according to its documented instructions (in particular on how the data set should be constituted with regard to data sources and categories), the scraper may be regarded as a processor. The AI developer will generally determine the instructions given in regard to the purposes and essential means of the scraping, and therefore be considered the controller.

19 In situations where an AI developer trains an AI model using a data set which has already been scraped by another entity (scraper) it is necessary to distinguish between the responsibility for the different data processing. The scraper and AI developer are, in principle, responsible for their own, separate, processing activities, since each of them determines the objectives and the essential means of its own processing. The scraper is not, in principle, responsible for the re-use of the data.

[2] EDPB Guidelines 7/2020 on the concept of controller, joint controllers and processor in the GDPR, adopted on 7 July 2021, paragraph 52: "The assessment of joint controllership should be carried out on a factual, rather than a formal, analysis of the actual influence on the purposes and means of the processing. All existing or envisaged arrangements should be checked against the factual circumstances regarding the relationship between the parties."

20 It is also possible that the scraper and the AI developer are jointly responsible for the processing instead. This could be the case when the scraper and the AI developer determine the purposes and means of the processing together (jointly).

> **Example 1:** Two organisations make a joint decision to develop an AI model, and agree that one of the organisations will perform the task to scrape data from the web to and create data sets for the development of the AI model, while the other organisation perform the development of the model, using the data sets created by the scraper. The collection criteria for the scraping of personal data are determined jointly by both parties. Even though they perform separate processing activities they determine the purposes and means of the processing (the web scraping and the development of the model) together and are jointly responsible for the processings of personal data.

3.2.2 Principles relating to processing of personal data

Lawfulness

21 Personal data should be processed lawfully, fairly and in a transparent manner in relation to the data subject.[3] When assessing whether the processing of personal data is lawful, the developer of the generative AI model as controller must consider the datasets (containing the scraped data) it intends to use, and/or the types of websites it would like to scrape data from.

22 Moreover, the purpose of the web scraping of personal data should be lawful.[4]

Purpose limitation

23 Personal data should be collected for specified, explicit and legitimate purposes, and not further processed in a manner that is incompatible with those purposes. The EDPB has provided further considerations on the purpose limitation principle in the context of the development of AI models in the Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models.[5]

Transparency

24 Personal data should be processed in a transparent manner in relation to the data subject.[6] This includes the obligation to inform the data subject about the processing of their personal data. However, when scraping large amounts of data from the internet, it is often difficult, impracticable or, even, objectively impossible to identify and inform the data subjects in an effective way. This makes compliance with the principle of transparency challenging.

25 Depending on how the data processing is precisely designed, it is possible that the controller of the web scraping might not have to inform data subjects individually about the web scraping, if providing the information proves impossible or requires disproportionate effort (Article 14 (5)(b) GDPR).

[3] Article 5(1)(a) GDPR.
[4] See Joint Statement on AI-Generated Imagery and the Protection of Privacy, issued on 23 February 2026, on the creation of non-consensual intimate imagery, defamatory depictions, and other harmful content featuring real individuals.
[5] EDPB Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, adopted on 17 December 2024, p. 64.
[6] Article 5(1)(a) GDPR.

26 Impossibility or disproportionate effort typically arises in circumstances where the personal data were not obtained from the data subject.[7]

27 Informing the data subjects individually may be impossible or require disproportionate effort in particular for processing "for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes", subject to the safeguards referred to in Article 89(1) GDPR.[8] While the word "in particular" clarifies that there may be other cases where it would be impossible or require disproportionate effort to inform data subjects,[9] this exception should not be *routinely* relied upon by controllers who are not processing personal data for the purposes of archiving in the public interest, for scientific or historical research purposes or statistical purposes.[10]

28 Where a controller seeks to rely on the exception in Article 14(5)(b) GDPR, it should assess the effort involved for the controller to provide the information to the data subject against the impact and effects on the data subject if they were not provided with the information.[11] The balancing exercise should be carried out with regard to the dataset as a whole and not on every single piece of personal data, and should take into account:

- the number of data subjects,
- the age of the data,
- any appropriate safeguards adopted.[12]

29 In the case of processing of personal data that do not require the identification of a data subject by the controller, the controller is not obliged to maintain, acquire or process additional information in order to identify the data subject for the sole purpose of complying with the requirements in the GDPR. In such cases, the controller is not required to comply with the rights of the data subject in Articles 15 to 20, unless the data subject provides additional information enabling their identification.[13]

30 In cases where informing data subjects individually proves impossible or would require disproportionate effort, the controller must take appropriate measures to protect the data subject's rights, freedoms and legitimate interests. One appropriate measure that controllers must always take, as specified in Article 14(5)(b), is to make the information publicly available.[14] This can be done for instance by publishing a privacy policy and/ or specific notice on the controller's website. This policy should include the information that is required by Article 14(1)-(2) GDPR, including the categories of personal data concerned, the purposes of the legal basis of the processing as well as a precise indication of the source(s) of the data. The information should indicate in particular whether the sources are publicly accessible or non-publicly available sources and, if the data is crawled and scraped from online sources, the information should include crawler's characteristics. In situations where it would be impossible or require disproportionate effort to provide a complete list of the

[7] Article 29 Working Party, WP260 rev.01, 29 November 2017, Guidelines on transparency under Regulation 2016/679 - endorsed by the EDPB, paragraph 62.
[8] Recital 62 GDPR.
[9] Recital 62 GDPR.
[10] Article 29 Working Party, WP260 rev.01, 29 November 2017, Guidelines on transparency under Regulation 2016/679 - endorsed by the EDPB, paragraph 61.
[11] Article 29 Working Party, WP260 rev.01, 29 November 2017, Guidelines on transparency under Regulation 2016/679 - endorsed by the EDPB, paragraph 64.
[12] Recital 62 GDPR.
[13] Article 11 GDPR.
[14] Article 29 Working Party, WP260 rev.01, 29 November 2017, Guidelines on transparency under Regulation 2016/679 - endorsed by the EDPB, paragraph 64.

sources, the sources should be included in the list to the greatest extent possible, and the list should be accompanied with an indication of the types of sources not included in the list, and the reasons for not including those sources.

31 Where possible, it is good practice to include information on the domain names and URLs of the scraped web pages, to provide them in a searchable format, and to include the date or period of collection.

32 In situations where an organisation obtains already scraped datasets to develop AI models, it is recommended to also provide the means to contact the controller from whom the data was obtained. A good practice is to link directly to the website of the controller for the web scraping, and to accompany the information with a clear, concise explanation of the conditions under which the personal data was collected. This information should be provided in clear and plain language.

33 Data subjects should also be informed about their rights under the GDPR, and on how they can exercise them. Information should be easy for users to find and understand. By doing this, the controller will not only meet GDPR requirements but also build trust with data subjects whose data were scraped.

34 Other appropriate measures to protect the data subject's rights, freedoms and interests, in addition to making the information publicly available, depending on the circumstances of the processing, may include: undertaking a data protection impact assessment and making it publicly available; applying anonymisation or pseudonymisation techniques to the personal data; minimising the data collected and the storage period; and implementing technical and organisational measures to ensure a high level of security.[15]

Example 2: In order to train the AI model, the AI developer collects a very large volume of data about data subjects from a variety of sources that extend back to 20 years ago, such as social media pages and discussion forums where no direct identifiers are included in the data. As the webpages might contain personal data of thousands or millions of individuals, the provider does not necessarily have the means to contact them. Given the fact that there is no direct link between the scraper and the data subject that enables the communications, that the publication dates of the data extend back to 20 years ago and that a very large dataset is collected, it would involve disproportionate effort on the part of the controller to provide all persons with the information required under Article 14 GDPR. The controller makes the required information publicly available on its web site that is easily accessible, in clear and plain language. The controller also applies appropriate measures, e.g. the exclusion of data from web pages where the data subjects are directly identifiable.

Example 3: A provider of an AI model intends to fine tune an LLM model that can produce coherent text completion based on a short input. To do so the provider collects personal data from a closed user group of 5 000 individuals, discussing in the last two years a specific topic and using a specific business or social jargon. The data subjects are directly identifiable through their full names, email addresses, and unique participant IDs. Given the fact that there are means to

[15] Article 29 Working Party, WP260 rev.01, 29 November 2017, Guidelines on transparency under Regulation 2016/679 - endorsed by the EDPB, paragraph 62.

contact the data subjects and that the data has been published or collected in the recent years, it would not involve disproportionate effort on the part of the controller to provide all persons with the information required under Article 14 GDPR. The controller should inform the data subjects individually about the scraping.

Data minimisation

35 The principle of data minimisation provides that personal data must be adequate, relevant and limited to what is necessary for the purposes for which they are processed.[16]

36 This means that the controller must only collect personal data necessary for the purpose of the processing. While there is a major challenge in scraping large amounts of data while limiting personal data to only what is necessary for the intended purpose of processing, the principle of data minimisation does not prohibit training an AI model or AI system with large volumes of data. Instead, the principle of data minimisation requires not to process personal data that would not be necessary, relevant or adequate. To comply with this, the controller should make an assessment before the scraping begins, so as not to collect personal data that is not necessary for the development of the model.

37 The controller should, before the collection of the data:

- consider using synthetic data instead of personal data,
- define precise collection criteria,
- undertake a data mapping and inventory process,
- **apply filters to exclude the collection of certain categories of data** where they are not necessary (e.g. bank transaction data, location/geolocation data, etc.) if they can be sorted out (e.g. if they rest in specific directories within file servers);
- **exclude from the collection certain categories of websites** (e.g. sites or social networks used mainly by minors) which structurally contain these personal data (e.g. data concerning vulnerable persons such as minors or certain sensitive data[17], such as financial information or location information);
- **exclude from the collection websites which clearly oppose the scraping** of their content, through the use of technical measures, such as the use of robots authentication to view content, robots.txt or ai.txt files, or CAPTCHA, which impose an action that can only be carried out by a human being and aim to prohibit access to pages by robots.

38 The controller should, during and after the collection of the data:

- **apply syntax-based filtering mechanisms (e.g. regular expressions)** during the collection to filter out personal data identifiable by their format (such as social security number, telephone number, etc);
- **where feasible, replace some or all real data with synthetic data** generated for testing, analysis, or training purposes, in order to reduce exposure to personal data;
- where feasible, anonymise or alternatively pseudonymise[18] personal data.

[16] Article 5(1)(c) GDPR.
[17] This term refers to data of a particularly sensitive nature and is distinct from special categories of data within the meaning of Article 9 GDPR. See also Article 29 Working Party Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679, wp248rev.01, endorsed by the EDPB, p. 9-10.
[18] EDPB Guidelines 01/2025 on Pseudonymisation, version for public consultation, adopted on 16 January 2025.

Accuracy

39 The principle of accuracy means that the personal data processed must be accurate and updated with regard to the specific purpose for which they are processed.[19] If scraping is used to process data obtained from different and/or unreliable sources, it will be difficult or even impossible to ensure the accuracy of the personal data. Complying with the principle of accuracy can also be challenging if the scraped personal data is stored for a long time and no longer up to date.

40 The principle of accuracy does not only apply to the collection and processing of personal data as such but also means that the controller should consider the final result of the AI training in order to minimize the risk of incorrect output. When the AI model, once it is placed on the market, is expected to produce (as output) data that constitute personal data, it should be compliant with the principle of accuracy.

41 Limiting the collection of personal data to what is necessary, it is crucial not only to comply with the principle of data minimisation, but also to make it easier for the controller to conduct the mapping of the personal data and to keep them updated and accurate.

42 To the greatest extent possible, it is recommended that the controller:

- **Scrape from reliable sources:** when possible, choose sources that are official, trustworthy, and maintained (e.g. a government registry, verified company profile) and avoid secondary or outdated aggregators.
- Timestamp the data: record when the data was scraped, to allow to show how up to date the data is.
- Validate the data **before using them in AI training:** correct formats (emails, phone numbers), spot-check random samples for factual accuracy.

3.2.3 Legal basis for the processing of personal data

43 Before carrying out scraping operations involving the collection of personal data or re-use of scraped data, the controller must identify a lawful basis to the processing.[20] The legal basis of legitimate interest is often relied on for scraping in the context of development of generative AI by private bodies. The other grounds referred to in Article 6(1) GDPR will generally be less likely to apply to web scraping in the context of development of generative AI by private bodies.

Consent

44 Consent would most probably not be an applicable legal basis, since it is often difficult to obtain the consent of individuals when personal data are collected indirectly and at large scale. Organisations scraping data (as 'third party data')[21] from the internet do not have a direct relationship with the data subject and are most probably not able to identify and obtain consent from each and every data subject before scraping data.

45 Sometimes data subjects are not even aware of the fact that their data are publicly available online. When data subjects make their personal data available online, for example on a web page accessible to everyone, this does not mean that the data subjects gave their consent to the scraping of their personal data for a specific purpose. In particular, the absence or

[19] Article 5(1)(d) GDPR.
[20] Article 6(1) GDPR.
[21] EDPB Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR, adopted on 8 October 2024; paragraph 7.

non-applicability of a robots.txt file on a web site does not amount to consent within the meaning of the GDPR.

Legitimate interest

46 Article 6(1)(f) GDPR provides a lawful basis for the processing of personal data where such a processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject, in particular where the data subject is a child.

47 In order for the application of Article 6(1)(f) to be lawful, three cumulative conditions must be fulfilled: the pursuit of a legitimate interest by the controller or by a third party; the need to process personal data for the purposes of the legitimate interests pursued; and that the interests or fundamental rights of the person concerned by the data protection do not take precedence over the legitimate interest of the controller or a third party (the so-called "balancing test").[22]

Condition 1: legitimate interest

48 The concept of interest is closely related to, but distinct from, the concept of the purpose of the processing.23 An interest is the broader stake or benefit that a controller or third party may have in engaging in a specific processing activity. While the GDPR and the CJEU recognized several interests as being legitimate,[24] the assessment of the legitimacy of a given interest should be the result of a case-by-case analysis.

49 As recalled by the EDPB in its Guidelines on legitimate interest, an interest may be regarded as legitimate if it is lawful, clearly and precisely articulated and real and present, not speculative.[25]

50 The EDPB has previously stated that the following examples may constitute a legitimate interest in the context of AI models: (i) developing the service of a conversational agent to assist users; (ii) developing an AI system to detect fraudulent content or behaviour; and (iii) improving threat detection in an information system.[26] With regard to the development and improvement of a general-purpose AI model, even when the precise use of the model has not yet been decided, it is recommended to refer to the objective pursued by the development of the model (indicating in particular whether it is commercial, public, scientific research, and whether it is internal or external to the organisation).

Condition 2: necessity

51 The second step of the assessment consists in determining whether the processing of personal data is necessary for the purpose of the legitimate interest(s) pursued. The

[22] EDPB Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR, adopted on 8 October 2024; Judgment of 17 June 2021, *M.I.C.M,* C-597/19, EU:C:2021:492, paragraph 106; Judgment of 4 May 2017, *Rīgas satiksme*, C-13/16, EU:C:2017:336, paragraph 28; *Bundeskartellamt*, paragraph 106. *Koninklijke Nederlandse Lawn Tennisbond vs Autoriteit Persoonsgegevens, C-621/22* , EU:C:2024:858, paragraph 37; *Mousse v Commission nationale de l'informatique et des libertés*, *SNCF Connect*, C-394/23 ; *HTB*, Joined Cases C-17/22 and C-18/22, EU:C:2024:738, paragraph 49.

[23] EDPB Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR, version 1.0, 8 October 2024, paragraph 14.

[24] EDPB Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR, version 1.0, 8 October 2024, paragraph 16.

[25] EDPB Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR, adopted on 8 October 2024, paragraph 17.

[26] EDPB Opinion on certain data protection aspects related to the processing of personal data in the context of AI models 28/2024, adopted on 17 December 2024, paragraph 69.

assessment entails two elements[27]: (i) whether the processing activity will allow the pursuit of the purpose; and (ii) whether there is an equally effective and less intrusive way of pursuing this purpose.[28]

52 When personal data is necessary for training generative AI, narrowing the collection criteria to exclude unnecessary collection of personal data, rather than scraping a wide part of the internet may be crucial to ensure the necessity condition is met. Using pseudonymised personal data, or synthetic data, may be another less intrusive way of pursuing this purpose.

Condition 3: balancing test

53 The balancing test requires, inter alia, the identification and description of the different opposing rights and interests at stake, i.e., on the one side, the interests, fundamental rights and freedoms of the data subjects; on the other side, the legitimate interests of the controller or a third party. The impact on the data subjects as well as their reasonable expectations are also relevant elements of the balancing.[29]

Data subjects' interests, fundamental rights and freedoms

54 In the context of web scraping for training a generative AI model, data subjects' interests may include, but are not limited to, the interest in self-determination and retaining control over their own personal data (i.e. the data that originally was published on a website for a certain purpose is then scraped and processed for generative AI training against the data subjects' wishes or without their knowledge).[30] Also, large-scale and indiscriminate data collection in the AI development phase may create a sense of surveillance for data subjects ('chilling effect'). This may lead individuals to self-censor themselves and risks undermining their freedom of expression. The 'chilling effect' on freedom of expression may be due to the possible use of an AI model to identify data subjects who intended to express themselves anonymously, especially if the AI model is connected to a web search tool, as well as to the possible use of the AI model for profiling within the meaning of the GDPR.[31] The generation of deep-fakes (pictures, videos, speech) also adds to the possible impact on data subjects' interests, rights and freedoms, as well as possibly on the right to dignity of the person concerned.[32]

Impact of the processing on the data subjects

<u>The nature of the data processed</u>

55 When scraping data from the internet, it should be recalled that - in addition to special categories of personal data and data relating to criminal convictions and offences that enjoy special protection under Article 9 and 10 of the GDPR respectively - certain types of personal data revealing highly private information, such as location data or financial data, may also be scraped. Personal data about minors may also be collected. The processing of such personal data may lead to significant consequences for data subjects and should be

[27] EDPB Opinion on certain data protection aspects related to the processing of personal data in the context of AI models 28/2024, adopted on 17 December 2024, paragraph 72.
[28] EDPB Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR, adopted on 8 October 2024, paragraph 29.
[29] EDPB Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR, adopted on 8 October 2024, paragraph 32.
[30] EDPB Opinion on certain data protection aspects related to the processing of personal data in the context of AI models 28/2024, adopted on 17 December 2024, paragraph 77.
[31] Article 4(4) GDPR.
[32] See also the section on Consequences for data subjects, at paragraphs 59-61 of these Guidelines.

considered when assessing the level of intrusiveness, notably in case of personal data relating to minors.

56 As a rule, the more sensitive or private the nature of the data to be processed, the more likely it is that the processing of such data will have a negative impact on the data subject, and the more weight should be attributed to it in the balancing test.

The context of the processing

57 The context of the processing and the specific data processing methods may influence the impact that the processing may have on the rights and interests of the data subject. The EDPB has previously stated that the use of web scraping in the development phase may lead - in the absence of sufficient safeguards - to significant impacts on individuals, due to the large volume of data collected, the large number of data subjects, and the indiscriminate collection of personal data.[33]

58 Since information published on the internet are rarely deleted, the scraped data may have been published during a long period of time. Scraping a larger amount of websites means the interference with the fundamental rights to privacy and to the protection of personal data increases.

Consequences for data subjects

59 The rights, freedoms and interests of the data subjects may be impacted by the envisaged processing. The greater the impact on the data subjects is, the less probable it is that the controller can rely on legitimate interest.

60 In practice, it might be difficult for data subjects to exercise their rights under the GDPR and to oppose the scraping of their data, notably when data are scraped by third parties. This means that, even if the data subjects are aware that their personal data will be scraped, it is difficult for them to make these data unreachable to scrapers. One of the main consequences of scraping in the context of generative AI is that given the current technical state of the art, once the model is trained, personal data cannot be easily deleted from a model.

61 The analysis of the possible further consequences of the processing should also consider the likelihood of these further consequences materializing. The assessment of such likelihood should be made taking into consideration the technical and organizational measures in place and the specific circumstances of the case.[34]

Expectations of data subjects

62 The data subjects' reasonable expectations play an important role when weighing their legitimate interest(s) and the interests or fundamental rights and freedoms of data subjects.

63 Reasonable expectations do not necessarily depend on the information provided to data subjects, as the mere fulfilment of the information obligations set out in Articles 12, 13 and 14 GDPR is not sufficient in itself to consider that the data subjects can reasonably expect a specific processing.[35] However, the information provided to data subjects may be

[33] EDPB Opinion on certain data protection aspects related to the processing of personal data in the context of AI models 28/2024, adopted on 17 December 2024, paragraph 86.
[34] EDPB Opinion on certain data protection aspects related to the processing of personal data in the context of AI models 28/2024, adopted on 17 December 2024, paragraph 90.
[35] EDPB Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR, adopted on 8 October 2024, paragraph 53.

considered when assessing whether data subjects can reasonably expect their personal data to be processed.[36]

64 Given the technological developments of recent years on generative AI, people may be aware that the data they publish online may be accessed, collected and reused by third parties. However, it cannot be considered that they can always expect such processing to take place in all situations, for all purposes, for all controllers' interests and for all types of data accessible online concerning them. Moreover, the impact of generative AI models, depending on the capability of the model, may also change in time. In particular, the controller should consider:

- the **nature of the source websites** (social networks, online forums, dataset dissemination sites, etc.)
- the **type of publication** and the **publicly accessible nature of the data** (for example, an article published on a freely accessible blog has a more public character compared to a post on a social network published with access restrictions). As a 'rule of thumb', when the data has been made public and visible to everyone by the data subject themselves, the data subjects are more likely to reasonably expect that their personal data may be processed by others.
- the **restrictions imposed by the scraped website**, for example through the implementation of technical measures such as the use of robots authentication to view content, robots.txt or ai.txt files, or the implementation of CAPTCHA, which are intended to impose an action that can only be carried out by a human being, and aim to prohibit access to the pages by robots. If data subjects are aware that a website implements these measures and a website is nevertheless scraped, it is less likely that they can expect the processing of their personal data by a scraping entity.
- the **nature and characteristics of the relationship between the data subject and the controller**: even though there is often no direct relationship between the data subject and the entity carrying out the web scraping or using scraped data, people may be aware that the data published online may be accessed, collected and reused by third parties;[37]
- the **characteristics of the data subjects** (minor, public figure, position of the data subject, etc.)

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. In this instance, data subjects can reasonably expect third parties to scrape that data to develop AI models.

Example 5: If a person uploads their data to a content sharing platform that prohibits scraping through the use of robots.txt files and the implementation of CAPTCHA, and expressly states on their site that they do not allow the use of their users' data for the development of AI models, they cannot reasonably expect third parties to scrape that data for that purpose.

[36] EDPB Opinion on certain data protection aspects related to the processing of personal data in the context of AI models 28/2024, adopted on 17 December 2024, paragraph 92.
[37] EDPB Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR, paragraph 54.

The role of mitigating measures

65 When for a specific processing the data subjects' interests, rights and freedoms override the legitimate interest(s) being pursued by the controller or a third party, the controller may consider introducing mitigating measures to limit the impact of the processing on data subjects. Mitigating measures are safeguards that should be tailored to the circumstances of the case. These additional measures should not be confused with measures that the controller is legally required to adopt anyway to ensure compliance with the GDPR.[38] If the data subject's interests, rights and freedoms override the legitimate interests being pursued, and sufficient mitigating measures are not taken, the processing cannot be based on Article 6(1)(f) GDPR.

66 Measures may include **technical or organisational** measures, such as:

a. Excluding data content from publications which might include personal data entailing risks for particular persons or groups of persons (e.g. individuals who might be subject to abuse, prejudice or even physical harm if information was released publicly).

b. Ensuring that certain data categories are not collected or that certain sources are excluded by default from data collection is important to comply with the principle of data minimization, but also to limit the impact of the scraping on data subject. This could include, for instance, certain websites that are particularly intrusive due to the sensitivity of their subject matter.

c. Limiting the collection by excluding scraping of certain personal data, such as:

   - personal data on sites, which are only accessible when logged into the site or service and are therefore not freely accessible, for example personal data published on a social network that requires login before getting access to that data;

   - personal data on sites that are freely accessible but in relation to which data subjects cannot be considered reasonably aware of the accessibility. For example, where a social network returns different outputs when the user agent of the scraping service is used and when the user agent of the web browser from the social media user is used, particularly where the scraping service receives more data (e.g. meta data such as timestamps or access statistics by using a technical API) than the regular user.[39] This could also be the case for an online directory or register in which anyone can freely look up an individual's name, address and contact details;

d. Imposing other relevant limits on collection, possibly including criteria based on time periods.

e. Establishing safeguards that contribute to increased transparency. For example, the controller can publish and make available, as widely as possible, information about data collection and data subjects' rights (e.g. through online articles), ensuring that an updated list of scraped websites is published.

[38] EDPB Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR, adopted on 8 October 2024, paragraph 57.

[39] For information about the term 'user agent', see for example User agent - Wikipedia, or RFC 9110 - HTTP Semantics.

f. Facilitating the exercise of individuals' rights, including putting in place a discretionary and prior right to object to strengthen individuals' control over their data. The controller can also create an opt-out list, which allows data subjects to object to the collection of their data on certain websites or online platforms by providing information that identifies them on those websites, including before the data collection occurs.

g. Deleting or anonymising the unnecessary personal data as soon as possible.

h. As referred to in the EDPB guidelines on pseudonymisation, pseudonymisation could be a measure to limit the direct identification of personal data. This could, in some cases, be done by replacing direct identifiers with placeholders and use these during the AI training.

i. Other additional measures that are relevant to protect personal data concerned by the further stages of the AI development phase, including measures to limit the risks of memorisation, regurgitation or attack of AI models or systems (e.g. deduplication).

These measures do not constitute an exhaustive list. The controller should make a case-by-case assessment to determine which measures are appropriate.

Example 6: An organisation that collects numerous publicly available online voice recordings, in order to develop and market a voice generation tool, without taking any additional measures to protect the training data or limit the risks of unlawful or malicious reuse, cannot rely on the legal basis of legitimate interest.

Example 7: An organisation aims to develop a generative AI system for text. It exclusively uses data from freely and publicly accessible online sources, where the data subjects have manifestly made the content public. It also excludes any content protected by copyright (i.e., using only content in the public domain or for which rights holders have not objected to text and data mining as permitted under Directive 2019/790 on copyright and related rights in the Digital Single Market). In addition, it implements a range of safeguards to limit data memorization and regurgitation, restricts problematic content generation through technical or contractual measures, facilitates the exercise of data subjects' rights when reidentification is possible, and clearly indicates data sources in a publicly available privacy policy. In such a case, the balancing test may generally be considered as met.

3.2.4 Special categories of personal data

67 Processing of special categories of data, such as data revealing racial or ethnic origin or political opinion, and concerning health or sexual orientation, is prohibited.[40] For the *intended* scraping of such data a derogation in Article 9(2) GDPR is required, in addition to a lawful basis under Article 6 GDPR.

[40] Article 9(1) GDPR.

68 Even though the controller should implement measures to prevent the collection of such personal data, it is not always possible to assess with reasonable certainty whether the scraping leads to the collection of special categories of personal data as defined in Article 9 GDPR until after the data has been scraped, thus when the processing activity has already been carried out.

69 It remains likely that, despite the organisational and technical measures, the controller residually processes special categories of personal data that it did not intend to collect and subsequently process. If such processing cannot be justified under EU data protection law, it will be unlawful for controllers to scrape such personal data in the context of AI development (subject to the considerations made below)

The case GC & Others, C-136/17

70 For the processing of personal data in the context of the activity of a search engine, the Court of Justice of the European Union ('the Court') has emphasized that the processing and dissemination of special categories of data, constitutes a particularly serious interference with the fundamental rights to privacy and the protection of personal data, and that there is no exemption from the prohibitions or restrictions contained in Article 9 or 10 GDPR for the activity of a search engine. However, the Court recognised that the specific features of the processing carried out by the operator of a search engine may have an effect on the extent of the operator's responsibility and obligations under those provisions,[41] and that the prohibition in Article 9(1) GDPR applies '*within the framework of his responsibilities, powers and capabilities*'.[42]

71 Against this background, the Court considered the prohibitions and restrictions of Articles 9 and 10 GDPR apply to the operator of a search engine only by reason of referencing (on a page with search results following a search on the basis of an individual's name) and via verification on the basis of a request by the data subject.[43]

The responsibilities, powers and capability test

72 The EDPB considers that the Court's ruling can be relevant for the incidental and residual collection of special categories of personal data in the context of web scraping for the training of an AI model.[44] The Court's ruling can be relevant only for processing activities similar to those in the judgement, and under similar circumstances, and should not be seen as a general exemption from the requirements in Articles 9 and 10 GDPR.

73 The controller should make a case-by case analysis to assess if the following conditions are met:

a. the processing activity has relevant similarities with the processing activity of a search engine, as referred to in the Court's judgement,

b. the processing includes only the incidental and residual processing of special categories of personal data, and no intentional processing,

[41] CJEU, Grand Chamber, Judgment of 24 September 2019, GC and Others C-136/17, paragraph 45.
[42] C-136/17, paragraph 48.
[43] C-136/17, paragraph 47.
[44] Incidental and residual collection of data should be understood as a processing of data that the controller has no intention of carrying out, but which will occur to some extent despite the implementation of measures to prevent such collection.

c. it is difficult/impossible to assess whether and to what extent the processing includes special categories of personal data, and therefore prevent the collection if such data, and

d. the controller implements measures '*within the framework of his responsibilities, powers and capabilities*', as mentioned in the Court's judgement, to prevent the dissemination of special categories of personal data.

The responsibilities, powers and capabilities when web scraping in the context of the development of generative AI models

74 If the **conditions a-c** are relevant for the specific processing activity, and the controller implements measures '*within the framework of his responsibilities, powers and capabilities*' to prevent the dissemination of special categories of personal data, the prohibition laid down in Article 9(1) GDPR will apply to a controller only within this framework of his responsibilities, powers and capabilities **(condition d)**. These responsibilities, powers and capabilities will include at least the following measures:

i. **Before the collection of the data** the controller should define precise criteria and apply filters to prevent the collection of special categories of personal data. The controller should also exclude certain categories of websites (e.g. sites or social networks mainly used by minors) which structurally contain these data (e.g. data concerning vulnerable such as minors or certain sensitive data).

ii. **After the collection of the data,** the controller should ensure that special categories of personal data that may have been collected despite the application of these measures is deleted from datasets immediately after the collection or as soon as it is identified. In particular, the controller should delete special categories of personal data upon request from the data subject if the request constitutes a plausible indication that the concerned data indeed fall under the prohibition laid down in Article 9(1) GDPR.

iii. **During the development of the AI model**, the controller should prevent extraction of special categories of personal data from the model and provide assurance regarding state-of-the-art resistance to privacy attacks. The controller should train and perform testing of the model and apply content (output) filters to the AI system to prevent the system from producing results and giving information to users that reveals special categories of personal data.

iv. **After the development of the AI model**, the controller using the model as part of an AI system should implement a process to constantly monitor the output generated by the AI system. When, despite the organisational and technical measures already implemented, special categories of personal data is included in the output, the controller should immediately take measures to prevent the production of such data (e.g. through updated or reinforced output filters, restricting certain prompts or functionalities). In the longer term, model unlearning or other processes with equal effect that might become available through technological progress should also be considered is necessary. If,

through technological progress, machine unlearning algorithms can guarantee deletion of specific data from a model this may provide an alternative to output filters or retraining the model.

75 The controller should be able to demonstrate, in line with the accountability principle, that the **conditions a-c** are relevant for the specific processing activity, and that the measures adopted *within the framework of his responsibilities, powers and capabilities*' **(condition d)** are relevant and effective to prevent the collection and the dissemination of special categories of personal data.

76 The implementation of these measures should be accompanied by regular verification and assessment - having regard to the development as well as to the use of the AI model - of the effectiveness of the measures implemented when performing web scraping activities. As a result of such verification and assessment, it is possible that additional or different organisational and technical measures should be implemented by the controller when the measures adopted are not effective.

77 The measures to be implemented should be in line with the state-of-the-art techniques available to limit the collection and other processing of special categories of personal data. It is however possible that in the future better privacy enhancing techniques to limit or even exclude the processing of special categories of personal data become available. Controllers should therefore constantly monitor technological developments to ensure that the state-of-the-art measures remain appropriate.

For the European Data Protection Board

The Chair

Anu Talus