epic.org

Electronic Privacy Information Center
1519 New Hampshire Avenue NW
Washington, DC 20036, USA

+1 202 483 1140
+1 202 483 1248
@epic.org
https://epic.org

COMMENTS OF THE ELECTRONIC PRIVACY INFORMATION CENTER

to the

EUROPEAN DATA PROTECTION BOARD

Regarding the

GUIDELINES 03/2026 ON WEB SCRAPING IN THE CONTEXT OF GENERATIVE AI

October 30, 2026

In July of 2026, the European Data Protection Board ("EDPB") issued a call for comments on their released Guidelines 03/2026 on Web Scraping in the Context of Generative AI (hereinafter the "Guidelines").[1] The Guidelines focus on situations where data is scraped by private entities from the internet for the purpose of training generative AI systems and address how this practice aligns—or may not align—with different General Data Protection Regulation ("GDPR") requirements and privacy principles. The Electronic Privacy Information Center ("EPIC") submits the following comments regarding the Guidelines.

EPIC is a public interest research center based in Washington, D.C., established in 1994 to focus public and regulatory attention on emerging privacy and human rights issues and to protect privacy, freedom of expression, and democratic values in the information age.[2] EPIC has long

---

[1] "Guidelines 03/2026 on web scraping in the context of generative AI," EDPB (July 8, 2026), https://www.edpb.europa.eu/public-consultations/guidelines-032026-on-web-scraping-in-the-context-of-generative-ai_en#no-back.

[2] EPIC, *About EPIC* (2019), https://epic.org/epic/about.html (last visited June 14, 2024).

engaged in international fora and comment opportunities to further protect privacy rights, particularly regarding new and emerging technologies.[3]

EPIC is thrilled that the EDPB is engaging on the complex and fast-moving issue of AI system development and training. Debates around appropriate and legal sources of quality training data have run rampant, hitting on privacy concerns, intellectual property issues, safety and security, exercise of individual rights, privileged and business-protected data, and more. While we applaud this initial effort, the Guidelines as currently structured fail to fully address both the scope of data scraping and concerns about the scraped data intended to be used.

First, the scope of the Guidelines leaves a significant practical gap in covering entities that engage in web scraping for AI training. Second, we remain unconvinced that web scraping for this purpose can be made compatible with the GDPR. Finally, the act of web scraping for AI training is antithetical to established privacy principles. We include at the end some recommendations that may address portions of the concerns raised, but ultimately conclude that web scraping for generative AI training is a violation of EU law.

The Guidelines' Scope Fails to Address Common Data Scraping and AI Training Practices

The Guidelines explicitly state at the outset that they apply to only two scenarios: i) an organization scrapes internet data or contracts a third party to scrape internet data for the purpose of

---

[3] "EPIC Comments on EU Metaverse Vision," EPIC (May 23, 2023), https://epic.org/epic-comments-on-eu-metaverse-vision/; "EPIC Comments on UK ICO Draft Guidance on Automated Decision-Making," EPIC (June 2, 2026), https://epic.org/epic-comments-on-uk-ico-draft-guidance-on-automated-decision-making/; "EPIC Comments to Dutch DPA on Emotion Recognition Prohibition Under the EU AI Act," EPIC (Dec. 17, 2024), https://epic.org/documents/epic-comments-to-dutch-dpa-on-emotion-recognition-prohibition-under-eu-ai-act/; "Comments of EPIC to the Office of the Privacy Commissioner of Canada Regarding the Update to Guidance on Handling Biometric Information," EPIC (Jan. 12, 2024), https://epic.org/documents/comments-of-epic-to-the-office-of-the-privacy-commissioner-of-canada-regarding-the-update-to-guidance-on-handling-biometric-information/.

generative AI development or ii) an organization obtains and uses a previously-scraped dataset from another organization.[4] Scoping the Guidelines to exclude entities that web scrape but do not train AI fundamentally defeats the purpose of the Guidelines. Focusing only on the AI companies developing the AI systems ignores data brokers—which perform an enormous amount of web scraping—by allowing them a simple loophole: they aren't scraping data EXCLUSIVELY for generative AI system training.

If data brokers are not included in the scope of these Guidelines, several of the requirements and recommendations are moot. The EDPB rightly highlights the importance of properly scoping the categories of data to be collected and respecting "do not scrape" signals and emphasizes that these steps must be undertaken prior to scraping. While AI developers can certainly review and clean data sets after they have been acquired, a good portion of harm is already done if the entities performing the scraping are not held to the Guidelines' standards.

The Guidelines as currently structured address only the actions of the AI companies once they have this scraped data, but ignore one of the established pipelines they obtain that data from, allowing scrapers to elide responsibility for violating individual privacy rights and putting individual safety at risk by scraping this data and offering it to AI developers. In order to properly address these risks and implement appropriate protections, the Guidelines must include any web scraping activity and must account for every entity in the web scraping supply chain. This includes entities that scrape data, middlemen that sell existing datasets to third parties, and the companies buying the data sets to ultimately train the AI systems at issue.

---

[4] Guidelines, Section 1, para 3.

The Guidelines Do Not Address Many Likely Violations of the GDPR

As noted in the definition of web scraping provided in the Guidelines,[5] web scraping is almost certain to include personal data. As such, the processing of that personal data—which includes its initial collection by web scraper— must comply with the requirements of the GDPR. Personal data includes any information "relating to an identified or identifiable natural person."[6] This includes both direct and indirect identification and would include inferences that can be used to identify an individual. The more personal data collected, the more likely the person is to be identified and increasingly sensitive and private information about that person may be revealed, particularly where automated systems combine data sets and make inferences from them.[7]

Before any personal data processing can occur (including the act of scraping it from the internet), a legal basis for that processing must be established. Article 6 of the GDPR sets forth the acceptable legal bases that may be used to process personal data: i) consent from the individual; ii) necessity to perform a contract; iii) compliance with legal obligation; iv) necessary for a vital interest of a person; v) necessary to perform a public interest task; or vi) legitimate interest.[8] The Guidelines identify legitimate interest as the most likely suited legal basis of the six to apply to web scraping for AI training. However, Recital 47 of the GDPR explicitly states that the interests of the data subject will likely override the data controller's interest where "personal data are processed in circumstances where data subjects do not reasonably expect further processing."[9] Since much of the personal data contemplated in scraping for training data was posted prior to AI training being a known risk, individuals are unlikely to have reasonably expected this risk. Further, the individuals

---

[5] Guidelines, Section 2, para 7.
[6] GDPR Art. 4(1).
[7] Daniel Solove, "The AI Inference Problem in Privacy Law," Solove on Tech (Oct. 21, 2026), https://danielsolove.substack.com/p/the-ai-inference-problem-in-privacy; Guidelines, Section 3.1, para 14.
[8] GDPR Art. 6.
[9] GDPR Rec. 47.

typically will have no established relationship with the scraper in question, giving them even less reason to anticipate that this separate company would access and use their personal data. The mere act of making information publicly available does not surrender any privacy interest the individual has in that data.[10]

Legitimate interest requires the satisfaction of three conditions, including a balancing test.[11] Those conditions are i) pursuit of a legitimate interest by the controller or third party; ii) processing personal data is necessary to pursue that legitimate interest; and iii) the interests or fundamental rights of the person involved do not override the legitimate interest of the controller. The first criterion can likely be established by most AI developers—they likely have a legitimate interest they can point to in developing the AI system. The second criterion may be more challenging if alternatives to scraped personal data (like using synthetic data) are available, but in some cases may still be satisfied. The third criterion—the balancing test—is where scraping for AI training is least likely to succeed. The Guidelines themselves note many of the individual interests, rights, and freedoms likely to be impacted by web scraping for AI training, including self-determination, control over personal data, a sense of surveillance and chilling effect on freedom of expression, risk of profiling, and deep-fake generation.[12] We would add that AI's aptitude for making inferences that constitute personal data (and sometimes very sensitive personal data)[13] is an additional risk and heightens the severity of risk for individuals.

---

[10] *See* Daniel Solove & Woodrow Hartzog, *The Great Scrape: The Clash Between Scraping and Privacy,* 113 California Law Review 1521, 1553 (2025), available at https://scholarship.law.bu.edu/cgi/viewcontent.cgi?article=4885&context=faculty_scholarship.

[11] Guidelines, Section 3.2.3, para 47; GDPR Art. 6(1)(f).

[12] Guidelines, Section 3.2.3, para 54.

[13] Daniel Solove, "The AI Inference Problem in Privacy Law," Solove on Tech (Oct. 21, 2026), https://danielsolove.substack.com/p/the-ai-inference-problem-in-privacy.

We appreciate the Guidelines' recommendations around mitigating measures for AI developers aiming to use legitimate interest as a legal basis. However, AI inferences and re-identification capabilities make it unlikely that the interests, rights, and freedoms of individuals can be mitigated in a way that would allow the controller interest to override them. Thus, we find it likely that most web scraping for AI training will not be able to properly establish a legal basis to do so and must therefore be disallowed.

Scraping Internet Data For Generative AI Training Does Not Align With Privacy Principles

The Guidelines attempt to parse how the privacy principles of the GDPR square with data scraping for AI training.[14] This is particularly important because several of the principles appear to be directly at odds with the practice of data scraping. We have examined each of the principles to determine whether it is possible for web scraping for AI training to be performed in such a way that will not violate those principles.

Lawfulness: The discussion of the lawfulness principle does not raise whether the act of scraping is, itself, lawful in instances where the data was not submitted to the website with scraping in mind. If information was posted with the understanding that it was for viewing—and web scraping for AI training was not envisioned (or addressed in the website's policy as a possible use of the data)—scraping may be unlawful. This is especially the case where the website's use policy or technical formatting forbids scraping or outside party use or replication of the data. We note that the Guidelines do address this problem in paragraph 45 under the discussion of consent, but it would be helpful to carry it through into this principle as well.

---

[14] GDPR Art. 5; Guidelines, Section 3.2.2.

Purpose Limitation: Regarding purpose limitation, the very act of scraping the personal data is likely a violation. Individuals may share personal data on a website for a specific purpose—promoting their work, networking, social engagement, making use of the website's services, etc. The initial action of making data available was for that initial purpose. Claiming that a separate company scraping that data to train AI models—an action that may not have even been a possibility at the time of posting—is consistent with the initial purpose is absurd.

Data Minimization: The Guidelines urges controllers to conduct an assessment prior to any scraping to ensure that personal data not necessary for model development is not included. The problem with this recommendation is that AI developers will argue that as much information as possible—including personal data—is necessary to create an AI model that functions best. While we certainly disagree with that claim, the larger point is that an assessment is not likely to result in meaningful engagement with or limitations on data to be collected. In paragraphs 37 and 38, the Guidelines provide very thoughtful recommendations around how to limit data collected, avoid the most sensitive personal data, and consider alternatives like synthetic data. We urge the EDPB to make this portion stronger, shifting from "should" to "must."

Accuracy: While the Guidelines include some recommendations regarding accuracy of data, they fail to note the most obvious and useful way to ensure training data is accurate: human review and curation. While the Guidelines recommend validating data prior to use in training, this merely includes correcting formats for consistency and "spot-check[ing] random samples for factual accuracy."[15] Random spot-checks in vast troves of data will solve nothing. Unless training data sets are comprised solely from trustworthy and up-to-date sources, an enormous amount of inaccurate, unverified, or outdated information will be included and will affect model development and

[15] Guidelines, Section 3.2.2, para 42.

performance. Further, individuals seeking to correct or delete inaccurate personal data that may be included in the data set will be without relief unless the entire model built from the inaccurate data is discarded and retrained with the updated information (as noted within the Guidelines).[16]

Transparency: The biggest and most significant challenge to scraping for AI training is the principle of transparency. The Guidelines note that notice to data subjects, as required under the transparency principle, may prove impossible or involve disproportionate effort.[17] This may well be true, but that is a problem for the entities seeking to benefit from the information to solve. When a practice is not compliant with the law, the burden is on the entity to either bring the practice into legal compliance (where possible) or, if the practice cannot be performed in a legally compliant way, to abandon that practice. The Guidelines raise as a possible option an exception that the very text of the GDPR notes is intended for "processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes."[18] While some AI developers may fall into these categories, the vast majority will not.

The Guidelines recognize that most AI companies will not fall under the commonly understood exception cases and stresses that the exception should not be routinely relied upon by controllers not processing data for those purposes.[19] It then raises a balancing test for companies still seeking to use the exception, weighing the effort required for the controller versus the effects on the data subject if they do not get this information.[20] However, the elements considered in this balancing test fail to factor in the immensely heightened risks to individuals due to the nature of AI. AI is

---

[16] Guidelines, Section 3.2.3, para 60.
[17] Guidelines, Section 3.2.2, para 24.
[18] GDPR Art. 14(5)(b); Guidelines, Section 3.2.2, para 25.
[19] Guidelines, Section 3.2.2, para 27.
[20] Guidelines, Section 3.2.2, para 28.

extremely adept at re-identifying individuals from "de-identified" data sets,[21] inferring increasingly sensitive information about individuals from seemingly low-risk data,[22] and using information about individuals to harass, dox, and manipulate those individuals.[23] Individuals can protect themselves from none of these harms if they are not informed that their information has been used by AI developers in the first place.

Example 2 in the Guidelines supposes that an AI developer scrapes huge volumes of data from many sources going back as far as 20 years. Since it would require "disproportionate effort" to notify the individuals affected, they put a notice on their website and exclude "directly identifiable data." This example is a legal fiction that pretends this low-effort action by the company makes their mass scraping and use of data acceptable. It does not. A notice on a web page does not provide the individuals with necessary information as we are all aware that, practically, affected individuals will never see this notice, making it "practically useless."[24] The very idea supposes that i) an individual is somehow aware that their data has been or may have been scraped for AI training (how they would

---

[21] Stefan Vamosi, Michael Platzer, and Thomas Reutterer, "AI-based Re-identification of Behavioral Clickstream Data," arXiv (Jan. 2022), available at https://ui.adsabs.harvard.edu/abs/2022arXiv220110351V/abstract; Oscar Thees, Roman Müller, and Matthias Templ, "Agentic AI-Powered Re-Identification: An Emerging, Scalable Threat to Mobility Microdata Privacy," arXiv (June 26, 2026), available at https://arxiv.org/html/2606.27936v1; "Reidentification risk for AI," IBM (May 22, 2026), https://www.ibm.com/docs/en/watsonx/saas?topic=atlas-reidentification.

[22] Daniel Solove, "The AI Inference Problem in Privacy Law," Solove on Tech (Oct. 21, 2026), https://danielsolove.substack.com/p/the-ai-inference-problem-in-privacy; "Gartner Predicts Most Privacy Incidents Will Stem from AI-Generated Inferences by 2029," Gartner (July 3, 2026), https://www.gartner.com/en/newsroom/press-releases/2026-07-30-gartner-predicts-most-privacy-incidents-will-stem-from-artificial-intelligence-generated-inferences-by-2029.

[23] Jean Want, Michelle Cera, and Blake Spendley, "Dox For Me, O Muse: Meta's New AI Agent Built Lists of People in Vulnerable Groups On Request," Hunterbrook (Sept. 28, 2026), https://hntrbrk.com/breaking-news/muse-doxxing; Sameer Hinduja, "An AI Bot Chose Violence," Cyberbullying Research Center (Feb. 24, 2026), https://cyberbullying.org/an-ai-bot-chose-violence.

[24] Daniel Solove & Woodrow Hartzog, *The Great Scrape: The Clash Between Scraping and Privacy,* 113 California Law Review 1521, 1553 (2025), available at https://scholarship.law.bu.edu/cgi/viewcontent.cgi?article=4885&context=faculty_scholarship.

become aware of this is unclear); ii) they somehow manage to guess or intuit which entity has conducted this scraping OR which entity is using the scraped data (again, unclear how they would learn this); and iii) they then are able to find the notice on this random developer's website that confirms their data has been taken and used for this purpose and they may have some rights around it. This entire scenario is preposterous.

Despite the exceptions and examples raised, the Guidelines appear to be conceding that the act of scraping data for AI training is not compatible with the requirements of the transparency principle. If this is the case (and we agree that it is), the appropriate and legally-mandated response is to declare that, unless a company can demonstrate an innovative approach to scraping that provides actual notice to individuals in a functional way, web scraped data cannot be used for AI training. Some practices by their nature cannot be done in compliance with privacy rights and law. Those practices are not legal.

Proposals for Solutions

While we strongly feel it would be best and most legally correct to forbid the use of web scraped data to train generative AI systems, we do have some proposals that may help protect individual rights and privacy in a more workable and practical way.

Centralized Database of Web Scrapers: We recommend that all companies scraping data, purchasing or otherwise acquiring data sets that include scraped data, and using those data sets for generative AI training register in a centralized and regularly maintained database with basic information such as:

- What data was scraped
- From what sources
- With what tool, including its name and how web services can identify it

- For what purpose
- How the data is cleaned and curated
- Any data sets the company has purchased or otherwise acquired that were originally scraped, and from whom they acquired these data sets
- How to opt out or exercise your privacy rights

This database would serve as: i) a centralized way for individuals to exercise their data protection rights that companies MUST respond to; ii) a useful way for law enforcement, journalists, public interest groups, and more to track and hold accountable organizations using this data; iii) a means to help websites effectively block scrapers that they do not want accessing their domain's data; and iv) a clear and consistent transparency exercise for organizations engaging with web scraped data. The European Union has recognized the usefulness of a centralized database in other contexts already. The AI Act requires high risk AI systems to register in such a database so that individuals engaging with an AI system know how their rights are being affected and can identify what company to enforce their rights with.[25]

In many instances, an individual is unlikely to know who captured their data or in some cases that their data was publicly available online at all. Creating a centralized location for individuals to find out where their data is likely to have been scraped, how it has been used, and an actual way to communicate with the company and opt out or request to have data corrected will go a long way to ensure that rights are actually protected and enforced. It will also make it much easier to identify companies that are not meeting their legal obligations.

An Opt-In System for Data Scraping: The default assumption should be that scraping a web page is not permitted for generative AI training unless a technical signifier, such as a robots.txt file,

---

[25] EU Artificial Intelligence Act, Art. 49.

indicates that the page is opting in to being scraped for that purpose. This shifts the burden of effort from the millions of online domains being targeted to the organizations seeking to profit from this data. A web page should not be forced to affirmatively enforce their rights by researching and implementing opt out mechanisms just because they are publicly available. Internet domain owners have many reasons to not want their data scraped for generative AI training: it may create liability for them depending on the context in which they obtained site data, it may make users less likely to engage or post, it may cause them traffic problems with bot queries, and more. This opt-in model aligns with the GDPR's consent model, creating informed and non-coerced consent to data processing. Furthermore, it aligns with the terms and conditions of many domains that already prohibit data scraping in their policies (that may be ignored by web crawlers or entities engaging in scraping).

Domains that affirmatively are willing to allow their web data to be scraped for generative AI training can create a technical signal opting in such as a robots.txt file. This signal could also be detailed enough to specify permitted purposes for scraping, such as only allowing scraping by public interest research institutions like universities for non-AI training purposes. This would allow for useful data sets to still be created while respecting the rights of both individuals and domain owners.

Cleaning Data Sets: While the Guidelines note that cleaning is a typical step in the web scraping process,[26] the description merely notes this as a technical exercise—removing duplications, cleaning up code and tags, and adjusting formatting for ease of readability by the AI system. We would encourage an update that requires other specific actions in cleaning the data sets. At a minimum, cleaning a data set should include:

- Human cleaning and review of the data

[26] Guidelines, Section 2, para 11.

- Removal of sensitive personal data
- Comprehensive data review for accuracy
- A documented risk assessment for any personal data that remains in the data set, including documentation of the legal basis for processing the data, why the data is necessary, whether the individual has been given notice, etc.

We also note that entities using scraped data must disclose the security measures and retention policy they have regarding this data publicly.

Conclusion

The AI industry has built norms of operation that lawmakers are struggling to address. We understand that challenging the actions of such a fast-paced and influential industry and forcing them to comply with legal mandates is a difficult undertaking. However, bad practices cannot be permitted simply because powerful industries want to continue them. The Guidelines have taken a thoughtful approach at trying to square an inherently privacy-violating practice with European law, but we remain unconvinced that this is possible. The Guidelines exclude major players in the data scraping world and should more firmly and clearly state that the default assumption is that web scraping for AI training is incompatible with the GDPR and privacy principles unless extraordinary measures or circumstances are in play. If the EDPB is not able to make such a statement, we would at the very least encourage consideration of some of the proposals listed above to allow individuals to protect their rights more easily in this challenging context. Privacy rights must come before industry profit.

Respectfully submitted,

Calli Schroeder
Calli Schroeder
Director, EPIC AI and Human Rights Program

Maria Villegas Bravo
Maria Villegas Bravo
EPIC Counsel