Center for AI and Digital Policy

Comments of

The Center for AI and Digital Policy

to the

European Data Protection Board

in response to the

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

The Center for AI and Digital Policy (CAIDP) welcomes the opportunity to respond to the European Data Protection Board's (EDPB) request for comments on the proposed guidelines on web scraping in the context of generative AI (Web Scraping Guidelines).[1] CAIDP commends the EPDB for providing an opportunity for the public to comment on the Web Scraping Guidelines and offers the following measures to further strengthen the Guidelines.

CAIDP Recommendations

1. Adopt genuine Privacy-Enhancing Technologies (PETs)
2. Strengthen lawful-basis safeguards to protect fundamental rights in web scraping
3. Expand ex-ante data protection impact assessments
4. Strengthen meaningful transparency and accountability through independent external audits
5. Strengthen remedy mechanisms

About CAIDP

CAIDP is the largest independent nonprofit AI policy research and education organization, with a global footprint across 130 countries and a network of more than 2,000 experts.[2] CAIDP has previously submitted a statement to EDPB on the use of facial recognition in law enforcement, recommending a ban on mass surveillance.[3] CAIDP has also advised the Dutch Data Protection Authority on Generative AI, calling for verifiable anonymization,[4] and on automated decision-making, urging mandatory fundamental rights impact assessments.[5] More

[1] The European Data Protection Board, *Guidelines 03/2026 on web scraping in the context of generative AI,* (8 Jul. 2026), https://www.edpb.europa.eu/public-consultations/guidelines-032026-on-web-scraping-in-the-context-of-generative-ai_en

[2] CAIDP, https://www.caidp.org/

[3] CAIDP, *Statement to European Data Protection Board on the Use of Facial Recognition Technology in the Area of Law Enforcement* (18 Jul. 2022), https://www.caidp.org/app/download/8405683363/CAIDP-Statement-EDPB-FRT-01082022.pdf?

[4] CAIDP, *CAIDP Advises Dutch Privacy Agency on Generative AI* (29 Jun. 2025), https://lnkd.in/p/gSn8VbMh

[5] CAIDP, *Dutch Data Protection Authority: Automated Decisions Require Real Transparency, Meaningful Contestability* (25 May 2026), https://lnkd.in/p/gAKdwxVP;

broadly, CAIDP has consistently engaged with the European Union, European Commission and other EU bodies on the drafting and implementation of the EU AI Act.[6]

CAIDP publishes the ***CAIDP AI Index***, which is the most comprehensive global assessment of national AI policies and practices measured against human rights and democratic values.[7] The ***CAIDP AI Index*** highlights the work of the EDPB, including its central role in clarifying how EU data protection law applies to AI.[8] CAIDP also considers whether national governments endorse certain AI resolutions of the Global Privacy Assembly in assessing alignment with democratic values.

1. Adopt genuine Privacy-Enhancing Technologies (PETs)

CAIDP urges the EDPB to strengthen the data minimization principle in the Web Scraping Guidelines by promoting the use of genuine Privacy Enhancing Technologies (PETs) precisely as it was originally conceived–techniques to "minimize or eliminate the collection of personal data."[9] PETs include utilizing low-data artificial intelligence, deleting unnecessary data, and creating techniques for robust anonymity.[10] PETs, where possible, minimize or eliminate the collection of personal data in the first instance and do not faciliate further transfers of personal data, which necessarily introduces the risk of date breach, misuse, and secondary use.[11]

[6] CAIDP, *Draft Guidelines for the classification of high-risk AI systems under the EU AI Act* (29 Aug. 2026), https://www.linkedin.com/posts/center-for-ai-and-digital-policy_caidp-eu-hi-risk-ai-26-july-2026-activity-7499240785456881664-nwmR.

[7] CAIDP, *CAIDP AI Index 2026*, https://www.caidp.org/reports/caidp-index-2026/

[8] Ibid, p.40

[9] *EDPB, Guidelines 03/2026 on web scraping in the context of generative AI, Version 1.0*, paras. 35-38; (*Madrid Privacy Declaration of 2009,* https://www.edps.europa.eu/sites/default/files/publication/09-11-03_madrid_privacy_declaration_en.pdf; Herbert Burkert, *Privacy-Enhancing Technologies: Typology, Critique, Vision in Technology and Privacy: the New Landscape* 143–67 (eds., Philip E. Agre & Marc Rotenberg 1997)

[10] CAIDP, Comments to the US OSTP on Advancing Privacy Enhancing Technologies (9 Jun. 2022), https://www.nitrd.gov/rfi/2022/87-fr-35250/Center-for-AI-and-Digital-Policy-PET-RFI-Response-2022.pdf

[11] M.Rotenberg, *Fair Information Practices and the Architecture of Privacy, (What Larry Doesn't Get),* 2001 Stan. Tech. L. Rev. 1 (2001); M. Rotenberg, *Protecting Human Dignity in the Digital Age*, UNESCO (Nov. 2000), https://unesdoc.unesco.org/ark:/48223/pf0000121984

The original conception of PETs supports compliance with Articles 5(1)(c), 25 and 32 GDPR,[12]which set forth obligations on data minimization, data protection by design and by default, and security.[13]

2. Strengthen lawful-basis safeguards to protect fundamental rights in web scraping

CAIDP urges the EDPB to strengthen safeguards for legitimate interest and ensure that the Web Scraping Guideline's framing does not encourage its overreliance as a lawful basis for processing personal data.[14] Web scraping of personal data pushes the limits of GDPR Art. 5 principles, such as "lawfulness, fairness and transparency," "data minimization," and "accountability," and raises fundamental rights concerns, including transparency, privacy, and personal data protection.[15] Consistent with Article 8 of the Charter of Fundamental Rights of the EU,[16] the EDPB should seek to ensure that these rights and principles remain central in its guidance for lawful basis, while fostering an innovation environment that incentivizes private entities to develop safer, rights-respecting technologies.

Individuals should not bear the burden of protecting their personal data. The suggestion in para. 64 that individuals may be aware that data they publish can be accessed, collected, and reused by third parties risks creates an unreasonable expectation that individuals can be responsible for protecting their own data obtained by others. The Global Privacy Assembly adopted resolution on the collection, use and disclosure of personal data to pre-train, train and fine-tune AI models emphasizes that the availability of such data does not automatically imply a lawful basis for its processing. [17] This would shift the burden of safeguarding personal data onto individuals and have a chilling effect on freedom of expression as the Web Scraping Guidelines acknowledge in para. 54.

The balancing test does not adequately protect individuals' rights and freedoms. Companies' commercial incentives and the power and information asymmetry between

[12] European Union, 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 (General Data Protection Regulation) Art.5(1), 25,32 (4 May. 2016), https://eurlex.europa.eu/eli/reg/2016/679/oj/eng

[13] Chris Buckley and Paul Mozur, *How China Uses High-Tech Surveillance to Subdue Minorities*, NY Times, (22 May 2019), https://www.nytimes.com/2019/05/22/world/asia/china-surveillance-xinjiang.html

[14] CAIDP, *CAIDP Advises UK on AI and Webscraping* (1 Mar. 2024), https://lnkd.in/p/gKS36Jsc

[15] Ibid

[16] Official Journal of the European Union, *Charter of Fundamental Rights of the European Union, 2000/C 364/01,* Art. 8, (18 Dec. 2000) *https://rm.coe.int/t-pd-2025-3-rev3-draft-guidelines-on-privacy-and-data-protection-in-th/48802bbb8b*

[17] Global Privacy Assembly, *Resolution on the collection, use and disclosure of personal data to pre-train, train and fine-tune AI models* (Sep. 2025), https://globalprivacyassembly.com/wp-content/uploads/2025/10/Resolution-on-the-collection-use-and-disclosure-of-personal-data-to-pre-train-train-and-fine-tune-AI-models.pdf

companies and individuals may tip the balancing test in favor of commercial interests. Thus, overreliance on legitimate interest risks prioritizing commercial gain over individuals' rights and freedoms and the public interest.[18]

Necessity must be demonstrated, not assumed. When assessing whether processing personal data is necessary, the Web Scraping Guidelines should direct controllers to prioritize the principle of data minimization under GDPR Art. 5 and consider less intrusive alternatives that do not rely on personal data.[19] The need for large-scale datasets to train generative AI does not establish that scraping personal data is necessary.

Additionally, CAIDP recommends that the Web Scraping Guidelines take into account the Court of Justice of the European Union ruling on the Passenger Name Record data, which holds that indiscriminate processing and retention of personal data may violate fundamental rights, even where data deletion obligations are observed.[20] Web scraping personal data to create datasets for Generative AI raises the same concerns.

3. Expand ex-ante data protection impact assessments

CAIDP urges the EDPB to strengthen the Web Scraping Guidelines by calling for **ex-ante** data protection impact assessments (DPIAs) ***before*** personal data are scraped or obtained for generative AI training, rather than treating them as possible safeguards.[21] The Guidelines should clarify and apply GDPR Art. 35 on data protection impact assessment, Art. 36 on prior consultation, and Art. 37 designation of a data protection officer.

Web scraping and processing personal data on a large scale to train GenAI models, is likely to result in a high risk to individuals' rights and freedoms.[22] DPIAs should establish whether personal data processing is necessary and proportionate, evaluate less intrusive and non-AI alternatives, and incorporate broader human rights impacts where public safety and civil rights are implicated.[23] Early assessment is essential to address harms, including re-identification, discrimination, and disclosure of personal data through model outputs, before

[18] CAIDP, *CAIDP Advises UK on AI and Webscraping* (1 Mar. 2024), https://lnkd.in/p/gKS36Jsc

[19] Ibid

[20] Court of Justice of the European Union (CJEU), *Ligue des droits humains v Conseil des ministres, Case C-817/19, Judgment of 21 June 2022*, ECLI:EU:C:2022:491, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:62019CJ0817

[21] EDPB, *Guidelines 03/2026 on web scraping in the context of generative AI*, Version 1.0, para. 34, (7 Jul. 2026), https://www.edpb.europa.eu/public-consultations/guidelines-032026-on-web-scraping-in-the-context-of-generative-ai_en

[22] Ibid, para 15

[23] CAIDP, *ICO Consultation on Generative AI - Lawful Basis for Web Scraping to Train Gen AI Models* (1 Mar. 2024), https://www.linkedin.com/posts/caidp-ico-ai-and-webscraping-mar-1-ugcPost-7169349418016559104-UmQT/

training makes effective remediation more difficult.[24] The Web Scraping Guidelines should also require documented findings, independent scrutiny, and reassessment following significant changes to a system's functionality or processing activities, consistent with the Global Privacy Assembly's emphasis on lifecycle assessments and external audits in its adopted resolution on Generative AI.[25]

4. Strengthen transparency and accountability through independent external audits

CAIDP recommends that the EDPB require independent external audits and data provenance documentation in the Web Scraping Guidelines to ensure traceability and accountability among controllers. [26] The Global Privacy Assembly resolution on Generative Artificial Intelligence Systems, sponsored by the European Data Protection Supervisor, states that developers, providers and deployers of generative AI systems should allow for external audits that can "independently assess how a model works, test outputs for inaccuracies and biases, and recommend effective measures to mitigate potential risks. The audits should be conducted across the AI lifecycle.[27] Data provenance documentation should include transparent and traceable sources, kinds, and scope of personal data used to train the model, along with clear definitions for the purposes for which the use of personal data has been consented.[28]

5. Strengthen remedy mechanisms

Given the high-risk web scraping poses to fundamental rights, CAIDP urges the EDPB to ensure that effective remedy mechanisms are in place to uphold data subjects' rights as established in GDPR Art 77-79. Paras. 29-33 in the Web Scraping Guidelines describe transparency and opt-out tools, but do not explain how individuals can exercise their rights, file a complaint, or how they can claim compensation. At the same time, the Web Scraping

---

[24] CAIDP, *ICO Consultation on Purpose Limitation in the Generative AI Lifecycle* (12 Apr. 2024), https://www.linkedin.com/posts/caidp-uk-ico-ai-purpose-specification-ugcPost-7184732868449816578-SBIY/

[25] Global Privacy Assembly, *Resolution on Generative Artificial Intelligence Systems*, 45th Closed Session (Oct. 2023),https://globalprivacyassembly.com/wp-content/uploads/2023/10/5.-Resolution-on-Generative-AI-Systems-101023.pdf; CAIDP, *Engineering Individual Rights into Generative AI Models: ICO Consultation - Generative AI Fourth Call for Evidence* (10 Jun. 2024), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6213178

[26] GPA, *Resolution on Generative Artificial Intelligence Systems* (Oct. 2023), https://www.google.com/url?q=https://globalprivacyassembly.com/wp-content/uploads/2023/10/5.-Resolution-on-Generative-AI-Systems-101023.pdf

[27] Department of Science, Innovation and Technology (DSIT), and AI Safety Institute, AI Seoul Summit, *International Scientific Report on the Safety of Advanced AI: Interim Report* (May 2024), https://assets.publishing.service.gov.uk/media/66474eab4f29e1d07fadca3d/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf

[28]GPA, *Resolution on the collection, use and disclosure of personal data to pre-train, train and fine-tune AI models* (Sep. 2025), https://globalprivacyassembly.com/wp-content/uploads/2025/10/Resolution-on-the-collection-use-and-disclosure-of-personal-data-to-pre-train-train-and-fine-tune-AI-models.pdf

Guidelines reference Article 11 GDPR, which risks placing the burden of identification onto individuals.[29]

CAIDP recommends that controllers provide accessible channels for data subjects to seek remedy and exercise their rights, and that supervisory authorities order the erasure of unlawfully scraped datasets and, where necessary, retraining of the model, as the EDPB indicated in the Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models.[30]

Closing Remarks

We thank you for considering our contributions. We welcome the opportunity to engage and discuss these recommendations in further detail.

Sincerely,



Merve Hickok
CAIDP President and Policy Director



Marc Rotenberg
CAIDP Founder and Executive Director



Ren Bin Lee Dixon
CAIDP Policy Group Associate



Tatjana Titareva
CAIDP Policy Group Member



Tatiana Zasheva
CAIDP Policy Group Member



Patrycja Niemczyk
CAIDP Policy Group Member

P.M.

Pritika Magima
CAIDP Policy Lead



Mert Özyiğit
CAIDP Policy Lead

[29] Ibid. Art. 11; paras. 29–33, 66(f) (7 Jul. 2026), https://www.edpb.europa.eu/public-consultations/guidelines-032026-on-web-scraping-in-the-context-of-generative-ai_en

[30] EDPB, *Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models* (17 Dec. 2024), https://www.edpb.europa.eu/documents/opinion-of-the-board-art-64/opinion-282024-on-certain-data-protection-aspects-related-to_en