AMERICAN BAR ASSOCIATION
Antitrust Law Section

321 North Clark Street
Chicago, IL 60654-7598
(312) 988-5550
antitrust@americanbar.org
www.ambar.org/antitrust

Attachment

October 30, 2026

Via: https://www.edpb.europa.eu/public-consultations/guidelines-032026-on-web-scraping-in-the-context-of-generative-ai_en#no-back

SUBJECT: Comments on the European Data Protection Board's Draft Guidelines Relating to Web Scraping in the Context of Generative AI.

Dear Sir/Madam:

On behalf of the American Bar Association Antitrust Law and International Law Sections, I respectfully submit this comment on the Joint Section Comment on the European Data Protection Board's Draft Guidelines Relating to Web Scraping in the Context of Generative AI.

The views expressed herein are being presented on behalf of the Antitrust Law and International Law Sections. They have not been reviewed or approved by the House of Delegates or the Board of Governors of the American Bar Association and, accordingly, should not be construed as representing the position of the Association.

If you have any questions after reviewing this report, we would be happy to provide further comments.

Sincerely,



Melanie L. Aitken
Chair, Antitrust Law Section



Edward Mullins
Chair, International Law Section

COMMENTS OF THE AMERICAN BAR ASSOCIATION ANTITRUST LAW SECTION AND INTERNATIONAL LAW SECTION ON THE CONSULTATION OF THE EUROPEAN DATA PROTECTION BOARD'S DRAFT GUIDELINES 03/2026 ON WEB SCRAPING IN THE CONTEXT OF GENERATIVE AI

October 30, 2026

The views expressed herein are being presented on behalf of the Antitrust Law Section and the International Law Section. They have not been reviewed or approved by the House of Delegates or the Board of Governors of the American Bar Association and, accordingly, should not be construed as representing the position of the Association.

The Antitrust Law Section and International Law Section of the American Bar Association (the Sections) welcome the opportunity to comment on the European Data Protection Board's (EDPB) Guidelines 03/2026 on Web Scraping in the Context of Generative AI (the Draft Guidelines).[1] The Sections appreciate the opportunity to provide these comments.

The Antitrust Law Section is the world's largest professional organization for antitrust and competition law, trade regulation, consumer protection, and data privacy as well as related aspects of economics. Section members, numbering over 12,000, come from all over the world and include attorneys and non-lawyers from private law firms, in-house counsel, non-profit organizations, consulting firms, federal and state government agencies, as well as judges, professors, and law students. The Antitrust Law Section provides a broad variety of programs and publications on all facets of antitrust/competition, consumer protection, and privacy. Numerous members of the Antitrust Law Section have extensive experience and expertise regarding similar laws of non-U.S. jurisdictions. For more than thirty years, the Antitrust Law Section has provided input to enforcement agencies around the world conducting consultations on topics within the Antitrust Law Section's scope of expertise.[2]

The International Law Section focuses on international legal issues, the promotion of the rule of law, and the provision of legal education, policy, publishing and practical assistance related to cross-border activity. Its members total over 11,000, including private practitioners, in-house counsel, attorneys in governmental and inter-government entities, and legal academics, and represent over 100 countries. The International Law Section's more than fifty substantive committees cover competition law, trade law, and data privacy and data security law worldwide

[1] Eur. Data Prot. Bd., Guidelines 03/2026 on Web Scraping in the Context of Generative AI (July 8, 2026), *available at* https://www.edpb.europa.eu/public-consultations/guidelines-032026-on-web-scraping-in-the-context-of-generative-ai_en#no-back [hereinafter Draft Guidelines].

[2] *See Antitrust Comments, Reports & Amicus Briefs*, AM. BAR ASS'N, ANTITRUST L. SECTION, http://ambar.org/atcomments.

as well as areas of law, which often intersect with these areas, such as regulatory compliance, mergers and acquisitions, and joint ventures. Throughout its century of existence, the International Law Section has provided input on debates relating to international legal policy[3] and has provided input on antitrust and competition law and policy for decades to authorities around the world.[4]

These comments reflect the expertise and experience of the Sections' members with data protection laws, consumer protection laws, and artificial intelligence (AI) regulations around the world. The Sections are available to provide additional comments, or otherwise assist the EDPB as it may deem appropriate.

The Sections commend the EDPB's proposal to undertake a thorough review of its laws and enforcement to protect consumers in the digital age, including through thoughtful engagement with consumers and other stakeholders alike.

SPECIFIC COMMENTS AND RECOMMENDATIONS

I. Current and Anticipated EU Legislation Support Generative AI Development

The Sections note that policies of the European Union (EU) encourage the development of AI capacity within the EU. As the EU AI Act sets forth, AI "contributes to a wide array of economic, environmental and societal benefits across the entire spectrum of industries and social activities."[5] The EU AI Act recognizes the need to protect fundamental rights and freedoms, and does so within a risk-based, uniform framework. Similarly, the GDPR states that "[t]he protection of natural persons in relation to the processing of personal data is a fundamental right."[6]

The Draft Guidelines try to balance all of these objectives and interests as relevant to Web scraping in the context of generative AI. The Sections believe that the Draft Guidelines for the most part strike a reasonable balance.[7] They acknowledge that technical and organizational measures, along with the fact-specific analysis of risks in the context of specific processing activities, have been recognized by the Court of Justice of the European Union (CJEU) as viable approaches to minimizing risk arising from the "incidental and residual collection by a controller

---

[3] *See About Section Policy*, AM. BAR ASS'N, INT'L L. SECTION, https://www.americanbar.org/groups/international_law/policy/about/.

[4] *See Blanket Authorities Initiatives*, AM. BAR ASS'N, INT'L L. SECTION, https://www.americanbar.org/groups/international_law/resources/policy/blanket-authorities/.

[5] Regulation (EU) 2024/1689 of the European Parliament and of the Council, of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), recital 4, http://data.europa.eu/eli/reg/2024/1689/oj.

[6] 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 (General Data Protection Regulation), recital 1, 2016 O.J. (L 119) 1 [hereinafter GDPR].

[7] *See* Draft Guidelines, *supra* note 1, ¶ 13.

of special categories of personal data."[8] The Sections encourage the EDPB to maintain, and even strengthen, the support suggested in this discussion.

II. The EDBP Should Consider Giving Fuller Attention to Emerging Alternative Market Practices Used for Data Acquisition

The Draft Guidelines address the scraping channel in detail and leave the competing channels to more general data protection legislation.[9] The EDPB recognizes that "web scraping can be a powerful tool for generative AI development, but it poses significant data protection risks."[10] Although Web scraping of public data has played an important part in pretraining large language models, the marginal return to additional scraped data has fallen sharply. The capability gains in frontier models are now predominantly attributed to proprietary data acquired from other channels, some of which enable the acquisition of data with lesser copyright or other legal and technical risks. Thus, scraping is but one of a number of means of acquiring data to improve an artificial intelligence model.

A developer facing a stringent necessity and balancing test for scraped data and comparatively unspecified regime for other economically valuable alternatives may shift data acquisition toward the alternatives. On the other hand, a relatively stringent test may also help incentivize greater attention and investment to develop more effective tools to ensure that effective technical and other protections are in place for any channel a developer wishes to use. The Sections recommend that the EDPB give full consideration to these emerging market practices, so that it protects and adequately incentivizes the protection of data subjects when collecting data to improve AI, regardless of the method of data acquisition, such as:

- *Commercial licensing from publishers and platforms*. Developers increasingly license content that was previously scraped through indiscriminate data collection, including news archives, image libraries, and user-generated content on social media. However, that content also may contain personal data, reaching the developer through a contract rather than a crawler.
- *Purchase of internal data from insolvent companies*. A market has emerged for internal communications and records of failed companies. The internal data in these transactions was never scraped but could raise similar privacy concerns from employees whose personal data may be transferred and processed without their consent.
- *Vertically integrated first-party corpora*. Several leading model developers are also affiliated with platforms that host large repositories of personal and social data. For these developers, the content that others must scrape is an internal resource governed by purpose limitation and the GDPR Compatibility Test.

---

[8] *See id.* at 3 (discussing Case C-136/17, GC & Others v. Comm'n Nat'le de l'Informatique et des Libertés (CNIL), ECLI:EU:C:2019:773 (Sept. 24, 2019)).

[9] *See* Draft Guidelines, *supra* note 1, ¶¶ 3-5.

[10] *See id.* ¶ 15.

- *Deployment-generated data*. User interactions with deployed models—prompts, feedback signals, and metadata—are among the most valuable proprietary inputs for model post training and personalization. Synthetic data can reduce dependence on human data subjects in verifiable domains such as computer code production but may be less relevant to models trained on text that reflects human judgment, preferences, and creativity.

The potential substitution problem has an immediate impact on the assessment of necessity and "less intrusive alternatives."[11] For example, content on an authentication-gated site held by an affiliated developer is likely to be excluded from scraping. It is possible that a substitute channel could be more intrusive than the scraping it replaces.

When assessing whether a less intrusive alternative exists, the Sections suggest that the final Guidelines direct the controller to compare the actual impact on data subjects irrespective of the technical method of data acquisition. The fact that an alternative source is obtained by contract, by acquisition of assets, or from within the controller's own organization, instead of Web scraping, does not automatically make it less intrusive. On the other hand, the EDPB should recognize and encourage the use of documented provenance and the practical ability to propagate objections and opt-outs, when controllers acquire data through channels where the counterparty has a direct relationship with the data subjects.

III. The Draft Guidelines Should Recognize the Practical Challenges of Filtering Special Categories of Data from Scraped Data

The use of public data from the Internet for the training of generative AI creates a difficult interaction with the rules on special categories of personal data (SCD). The Sections support the approach taken in the Draft Guidance to further advance the reconciliation of these concerns.

The GDPR's treatment of SCD reflects data that, by its nature, is "particularly sensitive" and so "merit[s] specific protection" given that its processing "could create significant risks to the fundamental rights and freedoms."[12] At the same time, SCD is somewhat broadly and imprecisely defined. Greater clarity and specificity around that term could both ensure that the data that is meant to and should be protected is protected, while providing more certainty that could be used to improve the tools and technical means used to protect privacy. At scale, data collection and sorting would benefit from tools that can both protect SCD and acquire more data at lesser risk. In more detail:

- *Public Internet data inevitably includes SCD*: As the EDPB recognizes, the public Internet that may be scraped includes many Websites likely to contain personal data.[13] The Internet empowers individuals to exercise their rights of freedom of expression

[11] *Id.* ¶ 52; *see also* Eur. Data Prot. Bd., Guidelines 1/2024 on Processing of Personal Data Based on Article 6(1)(f) GDPR, ¶ 29 (Oct. 8, 2024).

[12] GDPR recital 51.

[13] *See* Draft Guidelines, *supra* note 1, ¶ 7.

and information in countless different ways. As a result, it consists of an enormously broad and heterogenous mixture of information, including individuals discussing their families, political opinions, hopes and beliefs. As a result, public Internet data is replete with SCD.

- *Protecting SCD is critically important:* The EDPB observes in several places that serious harm can come from scraping SCD from the public Internet. Harm includes privacy violations, threats to the safety and well-being of minors, and even physical injury or abuse.[14] Artificial intelligence also may memorize SCD, enabling inferences to be drawn about an individual when combined with other data[15] or, in the most serious cases, raise concerns about surveillance and chill individual expression.[16] Data subjects may not even know their data is being scraped and, therefore, will not assert any available rights to protect themselves.
- *It is not currently feasible to completely eliminate all SCD from training data at scale*: The Sections support the general approach in the Draft Guidelines of not collecting public data from sensitive Websites and taking reasonable measures to try and filter out SCD, as discussed above.[17] However, completely removing SCD from training data at the scale of most scraping is not yet currently feasible. Compounding this challenge is uncertainty around the definition of SCD, which can complicate the development of effective tools and best practices. While protecting against SCD risks will inherently involve interpretation and some uncertainty, better clarity would help protect and balance the goals of the Draft Guidelines. That is particularly so because the definition of SCD includes not only information that prima facie reveals or concerns SCD, but also information that could indirectly disclose SCD. Given the challenges of making this assessment at scale, clearer guidance would make it easier to strike a better balance that meets more of the EDPB's goals.

Nonetheless, the Sections recommend that the EDPB revise the Draft Guidelines to indicate that the heightened scrutiny and balancing approach articulated in *GC & Others*[18] is not the exclusive framework for assessing the legality of SCD processing. To the extent that scraping is employed, the Draft Guidelines suggest that model providers must work to filter out all SCD from their output.[19] That is currently very difficult to achieve (see above). The final Guidelines might therefore provide further guidance on other situations in which SCD could be processed in model development. This should include when SCD can be processed on the basis it has been "manifestly made public by the data subject,"[20] as would be the case with, for example, the political affiliations of President Biden and European leaders.

---

14 *See id.* ¶¶ 55-56.

15 *See id.* ¶ 14.

16 *See id.* ¶ 54.

17 *See id.* ¶¶ 74, 76-77.

18 *See supra* note 8.

19 Draft Guidelines ¶ 74(iv).

20 GDPR art. 9(2)(e).

* * *

The Sections appreciate the opportunity to comment and remain available to respond to any questions about these comments or provide additional assistance to the EDPB in connection with its consultation.