BSA SUBMISSION ON THE EUROPEAN DATA PROTECTION BOARD DRAFT GUIDELINES ON WEB SCRAPING IN THE CONTEXT OF GENERATIVE AI

October 30, 2026

The Business Software Alliance (www.bsa.org) welcomes the opportunity to provide input on the European Data Protection Board's (EDPB) Draft Guidelines on Web Scraping in the Context of Generative AI.

BSA is the global trade association of the enterprise software industry, representing companies[1] that are leaders in artificial intelligence, cybersecurity, cloud computing, quantum, and other breakthrough technologies. We work in more than 20 markets in the US, Europe, and Asia, advocating for policies that build trust in technology so that every industry sector and the public can benefit from innovation. BSA's members provide privacy- and security-protective services to business customers across every sector of the economy.

BSA supports the EDPB's effort to explain how the General Data Protection Regulation (GDPR) applies when organisations process personal data in connection with web scraping in the context of generative AI. The Draft Guidelines appropriately recognise several important points, including that determining an organisation's role requires a fact-specific analysis, that consent will generally not provide a workable legal basis for large-scale web scraping, that legitimate interests may support processing when the requirements of Article 6(1)(f) are met, and that data minimisation does not prohibit training an AI model with large volumes of data. BSA recommends that the final Guidelines build on this foundation by:

- Retaining and strengthening the Draft Guidelines' fact-specific approach to determining whether an organisation acts as a controller, joint controller, or processor.
- Preserving the Draft Guidelines' recognition that consent is generally impractical for large-scale web scraping and can have important downsides.
- Continuing to support the use of legitimate interests for web scraping-related processing, including by assessing the purpose ultimately served by a model or system and ensuring necessity and balancing tests account for technical feasibility and realistic alternatives.
- Emphasizing that data minimisation is a purpose-specific and lifecycle-based principle. The final Guidelines should continue to emphasize that large and diverse datasets may be necessary for some AI systems, while giving sufficient recognition to downstream measures such as filtering,

[1] BSA's members include: Adobe, Alteryx, Amadeus, Asana, Atlassian, Autodesk, Avalara, Bentley Systems, Box, Cisco, Cloudflare, Cohere, Cohesity, Dassault Systemes, Databricks, Datadog, Docusign, Dropbox, Elastic, EY, Graphisoft, HubSpot, IBM, Kyndryl, MathWorks, Microsoft, Notion, Okta, OpenAI, Oracle, PagerDuty, Palo Alto Networks, PTC, Rubrik, Salesforce, SAP, ServiceNow, Shopify Inc., Siemens Industry Software Inc., TrendAI, TriNet, Veeam, Workday, Zendesk, and Zoom Communications Inc.

pseudonymisation, anonymisation, deletion, and output-level controls, which may be applied at different stages of the AI lifecycle.

- Ensuring that transparency requirements do not require disclosure of information that may undermine the privacy of data subjects.

1. The Final Guidelines Should Continue to Support a Fact-Based Approach to the Roles of Controllers and Processors.

BSA supports Paragraphs 16–20 of the Draft Guidelines, which recognise that determining whether an entity acts as a controller, joint controller, or processor requires a case-by-case assessment of the relevant facts and processing activities. The Draft Guidelines also correctly state that the entity performing the scraping is not necessarily the controller and may act as a processor when it conducts scraping on behalf of an AI developer and follows the developer's documented instructions.

This fact-specific approach to assigning roles under the GDPR is important for enterprise technology providers.[2] An organisation may supply infrastructure, tools, data-management services, or other technical capabilities to a customer without deciding why the customer processes personal data. It may also exercise discretion over nonessential technical or organisational details needed to provide a secure and reliable service. Those activities do not, by themselves, cause the provider to be treated as a controller or joint controller.

The final Guidelines should build on this analysis by emphasizing two points:

- First, analysing whether an entity acts as a controller or a processor should be conducted separately for each processing operation. For example, an organisation may play one role for collection but another for dataset preparation, model training, or fine-tuning.
- Second, a processor may make reasonable choices about implementation of a controller's instructions, such as technical and security measures, without altering its status as a processor. This is consistent with the EDPB's recognition that a controller may leave decisions about "nonessential means" of processing to the processor.[3]

2. The Final Guidelines Should Preserve the Recognition That Consent Is Often Impractical.

BSA also supports the Draft Guidelines' recognition that consent is often not the appropriate legal basis for large-scale web scraping. As the Draft Guidelines explain, organisations ordinarily lack a direct relationship with, and may be unable to identify, every data subject before collection of data through web scraping. The

[2] *See* BSA, Controllers and Processors: A Longstanding Distinction in Privacy (July 30, 2026), https://www.bsa.org/policy-filings/controllers-and-processors-a-longstanding-distinction-in-privacy.
[3] *See* EDPB, Guidelines 07/2020 on the Concepts of Controller and Processor in the GDPR, version 2.1 (July 7, 2021), https://www.edpb.europa.eu/system/files/documents/2023-10/EDPB_guidelines_202007_controllerprocessor_final_en.pdf.

final Guidelines should explain that this is not merely an administrative difficulty. In some contexts, requiring consent can impair privacy, safety, and fairness outcomes.

A cybersecurity system trained to identify malicious activity, for example, cannot depend on bad actors consenting to the use of relevant information. Consent requirements may also produce datasets that are less representative if participation varies significantly among communities. That can reduce the accuracy or reliability of an AI system and increase the risk of skewed outcomes. Legitimate interests, by contrast, require the controller to identify the concrete purpose, assess necessity, weigh effects on individuals, and adopt appropriate safeguards. That analysis can produce meaningful protection without imposing an infeasible consent requirement.

3. The Final Guidelines Should Provide More Practical Guidance on Using Legitimate Interests as a Workable Legal Basis.

BSA supports the Draft Guidelines' recognition that the basis of legitimate interests under GDPR Article 6(1)(f) is often relied on for web scraping to develop generative AI and that other Article 6 bases will generally be less likely to apply. To build on this, BSA recommends the final Guidelines:

- State directly that legitimate interests is an appropriate and workable legal basis for this processing when a controller satisfies the established three-part test.
- Provide practical guidance on applying each element of that test, which requires (1) the pursuit of a legitimate interest by the controller or a third party, (2) the need to process personal data for the purposes of the legitimate interest pursued, and (3) a balancing test assessing whether the data subject's interests or fundamental rights take precedence over the legitimate interests of the controller or third party.

We recommend clarifying how each part of this test applies, in practice.

a. *First condition: The Legitimate Interest Should Be Defined as the Purpose of an AI Model or System, Not Simply "AI Training."*

Companies train AI systems to achieve many purposes, including detecting cybersecurity threats, preventing fraud, identifying harmful content, improving accessibility, assisting users, and supporting scientific research. The Draft Guidelines already reflect this by recognising interests such as fraud detection and threat detection and by recommending that developers of general-purpose models identify their objective for a model's development. Organisations also increasingly use AI to pursue legitimate interests they already pursue through other tools.

BSA recommends the final Guidelines:

- Recognise that the relevant legitimate interest is the ultimate purpose the AI model or system is intended to achieve, not simply "AI training."
- Confirm that the use of AI does not, standing alone, displace a legitimate interest that would support comparable processing by other means.

- Clarify that when an organisation adopts AI to pursue an existing objective, it should apply the Article 6(1)(f) test to the new processing, including any new risks and safeguards, without any presumption that the legal basis changes merely because the technical tool has changed.

For example, an organisation that already processes publicly accessible information to identify malicious cyber activity should be able to rely on the same legitimate interest if it adopts an AI system for that objective, provided it assesses the new processing appropriately.

b. Second condition: Clarify that a Less Intrusive Alternative Must be Reasonably Available and Equally Effective.

BSA agrees with the Draft Guidelines' necessity analysis, which asks whether the processing will allow pursuit of the relevant purpose and whether an equally effective, less intrusive alternative exists.

BSA recommends that the final Guidelines provide practical ways to implement this analysis, including:

- Confirming that an alternative qualifies only if it is reasonably available, technically feasible, and comparably effective for the identified purpose.
- Specify that a method is not an equally effective alternative if it:
    - cannot produce a sufficiently representative dataset;
    - cannot achieve the required accuracy or security objective; or
    - would impose disproportionate technical constraints.
- Present synthetic and pseudonymised data as context-dependent options, not as preferred substitutes that controllers must disprove case by case. These techniques can be valuable safeguards, but they are not interchangeable with real-world data for every model, purpose, or development stage. Synthetic data may reproduce limitations in the data or assumptions used to generate it, and pseudonymisation may reduce identifiability without preserving every feature a particular use requires. If these options are presented as default requirements, rather than illustrative options, it could undermine legitimate AI training in circumstances where fully synthetic data or full anonymisation would materially degrade the quality of the AI model being trained.
- Recognise that more targeted collection is not necessarily less intrusive in every respect or equally effective, since a narrower dataset may omit relevant languages, regions, user groups, rare events, or emerging threats.
- Permit controllers to consider these consequences when documenting why particular categories and volumes of data are necessary.

c. Third condition: Confirm No Single Factor or Safeguard Determines the Outcome of the Balancing Test Across All Web Scraping Activities.

The Draft Guidelines correctly state that the balancing test depends on the nature of the data, context, likely consequences, reasonable expectations, and technical and organisational measures. They also provide a non-exhaustive list of potential safeguards.

BSA supports this case-specific structure and recommends that the final Guidelines:

- Expressly state that no single factor or safeguard determines the outcome of the balancing test across all web scraping activities.
- Clarify that technical signals such as robots.txt may provide relevant context but are not dispositive evidence of an individual data subject's expectations, nor a privacy-law prohibition in every case. These signals:
    - may be implemented by website operators for reasons unrelated to personal data; and
    - may apply across content supplied by many organizations.
- Direct controllers to weigh these signals together with other factors, including the source and accessibility of the data, the relationship among the parties, the type of content, the purpose of the processing, and the safeguards applied.
- Recognise that safeguards can operate at multiple points in the AI lifecycle, including defining collection criteria; excluding sources or data categories that present heightened risks; deduplicating data; filtering direct identifiers; limiting retention; applying pseudonymisation or anonymisation where feasible; providing public information and accessible rights mechanisms; testing for memorisation; and reducing the likelihood that a deployed system will reproduce personal data.
- Avoid suggesting that organizations create individual opt-out lists as an effective way to implement opt outs. Paragraph 66(f) suggests that controllers can "create an opt out list" to allow individuals to object to the collection of their data on certain websites or platforms. However, even if one organization could create such a list for a single website, this one-by-one list creation does not work at scale. This approach should not be presented as an effective option in the final Guidelines.
- Confirm that the appropriate combination of safeguards depends on the processing, and that a controller need only show that its measures reduce the identified risks, not implement every illustrative measure.

4. The Final Guidelines Should Apply Data Minimisation Principles Practically Across The AI Lifecycle.

BSA supports Paragraph 36 of the Draft Guidelines, which expressly recognises that data minimisation does not prohibit training an AI model or system with large volumes of data. This clarification should appear prominently in the final Guidelines. Data minimisation requires personal data to be adequate, relevant, and limited to what is necessary for the specified purpose; it is not a numerical cap on dataset size.

For many AI systems, dataset breadth is relevant to whether the model can recognise varied language, contexts, and rare events and whether it performs consistently for different groups. A controller should therefore be able to demonstrate why the volume and diversity of data are relevant to the intended purpose. The analysis should distinguish the amount of data from the amount of personal data and should focus on whether particular categories or fields are necessary, not whether the overall dataset is large.

The final Guidelines should also recognise that minimisation measures may be most effective at different stages. Some unnecessary sources or categories can be excluded before collection. Other personal data cannot reliably be identified until after extraction and may be addressed through cleaning, filtering,

pseudonymisation, deletion, or retention limits. Still other risks, including memorisation or regurgitation, may be better addressed through training methods, testing, and output controls. A lifecycle approach allows controllers to use the measure that most effectively reduces the particular risk.

BSA's recent comments on scientific research similarly urged the EDPB to avoid an overly rigid hierarchy among anonymised, pseudonymised, and identifiable data and instead support proportionate, risk-based approaches that reflect the realities of data-driven research. The final Guidelines should apply the same principle here. Anonymisation and synthetic data can be valuable, but requiring them when they would defeat the purpose of the processing, materially reduce utility, or are not technically feasible would not provide a workable application of the data minimisation principle.[4]

5. The Final Guidelines Should Not Require Disclosure of Information That May Undermine Privacy of Data Subjects.

The Draft Guidelines appropriately recognise that transparency requirements may be impossible or difficult to apply in an effective way in the context of certain web scraping activities. As the Draft Guidelines recognise, GDPR Article 14(5)(b) sets out a way for companies to support transparency when notices to individual data subjects are impossible or involve disproportionate effort — by making information about that processing public.

We are concerned the Draft Guidelines apply this requirement in a manner that may require the disclosure of information that would undermine the privacy of individual data subjects. For example, Paragraph 30 states that such public disclosures should include the categories of personal data concerned, the purpose of the legal basis of processing, and the "precise indication of the source(s) of the data." It goes on to say that if it is impossible to provide a complete list of sources, the sources should be listed to the greatest extent possible and reasons for not including other sources should be identified. In addition, the Draft Guidelines state that 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 period of collection."

These requirements appear to go beyond the GDPR and can undermine individual privacy. For example, a searchable list of URLs for training data enables re-identification of data that may have been subject to privacy-protective measures. It also allows bad actors to link individuals to specific webpages from which their data was obtained, ultimately enabling the collection of more information about individuals.

In addition, the Draft Guidelines suggest that additional transparency measures may include undertaking a data protection impact assessment (DPIA) and making it publicly available. We strongly recommend against suggesting that DPIAs be made public, which both goes beyond the GDPR's requirements for DPIAs and undermines their purpose. DPIAs are used as a robust internal accountability tool for organizations to identify risks arising from their processing; if organizations are encouraged to make such assessments public, it

[4] BSA, Submission on the European Data Protection Board's Draft Guidelines on Scientific Research (June 18, 2026), https://www.bsa.org/policy-filings/europe-bsa-response-to-the-edpb-draft-guidelines-on-scientific-research.

changes the incentives for conducting robust assessments. A DPIA for model development also raises more specific concerns, because it would typically include commercially sensitive measures like the technical and organizational security practices adopted in the model and the precise data sources relevant to its training.

The Final Guidelines should:

- State that providing a searchable list of URLs is an illustrative example, not a de facto requirement.
- Recognise category-level description of sources as sufficient for scenarios in which providing a list of specific sources creates risks to privacy or security.
- Avoid suggesting that DPIAs be made publicly available.

Conclusion

The Draft Guidelines take several important steps toward a workable application of the GDPR to web scraping for generative AI. BSA supports their recognition that roles must be determined case by case, consent is often impractical, legitimate interests can support processing, and data minimisation permits the use of large datasets when justified by the purpose. The final Guidelines should provide more practical guidance on these principles.

BSA appreciates the opportunity to provide these comments and looks forward to continued engagement with the EDPB.

For further information, please contact Thomas Boué,

Vice President and Director General, Policy – EMEA, at thomas@bsa.org.