Public Consultation Response to EDPB Guidelines 03/2026 on Web Scraping in the Context of Generative AI (v1.0)

To: European Data Protection Board, via the public consultation on Guidelines 03/2026 on web scraping in the context of generative AI

From: The Synthetic Outlaw Observatory Foundation (a 501(c)(3) nonprofit organisation)

Author: Jonathan Gropper, JD, Fulbright Specialist in AI governance; author of *The Synthetic Outlaw: How AI Defeats the Rules Without Breaking Them*

Date: 1 October 2026

Re: Comments on Guidelines 03/2026 on web scraping in the context of generative AI, version 1.0, adopted 7 July 2026 (consultation closes 30 October 2026)

We respectfully submit these comments on the draft guidelines. The Board has posed no specific consultation questions, so we address the operative provisions directly, with paragraph references to the adopted draft. Each paragraph number and quotation below was verified against the adopted v1.0 text.

The Synthetic Outlaw Observatory Foundation is a 501(c)(3) nonprofit that scores and publishes AI governance incidents. The dataset and the limits of our evidence are described in the Appendix. I coined the term synthetic outlawry for systems that defeat a rule's purpose while remaining formally compliant.

Executive summary

The draft guidelines are strong on paper. Our contribution is the enforcement gap between documented controls and demonstrated effectiveness. Three gaps. Three fixes.

We have no scored incident of an AI training scraper defeating robots.txt or CAPTCHA specifically. Our argument extrapolates from demonstrated agentic capabilities documented in the dataset (see Appendix). The draft itself documents

the relevant capability class at paragraphs 9 and 66(c), and we rely on that grounding alongside our data.

First, opt-out mechanisms and technical measures. Robots.txt. ai.txt. CAPTCHA. Crawler identification. The draft treats honoring them as compliance-relevant, at paragraphs 37, 64, and 66(f). Agentic scraping pipelines can rotate user agents and IP addresses, render JavaScript, resolve CAPTCHAs, and route through intermediaries. A controller can document respect for every signal while its pipeline defeats them. Compliance on paper. Circumvention in practice.

Second, legitimate interest assessments and mitigation documentation that are formally complete but never tested against what AI scrapers can actually do. Paragraph 61 assesses the likelihood of further consequences by reference to "the technical and organizational measures in place." We submit that a measure defeated by current capabilities should not be treated as in place for the purposes of that assessment.

Third, provenance depth on the acquirer side of brokered datasets. Paragraph 4's broker exclusion is a scope choice, not a legal vacuum: brokers remain controllers of their own collection, and paragraph 19 makes the acquirer responsible for its own processing. But nothing in the draft obliges the acquirer to establish the provenance of the dataset it buys. The gap is the depth of the acquirer's provenance obligation, not a responsibility vacuum.

Our proposals are concrete and in scope. They change the operative text, not the architecture of the guidelines:

1. Require controllers to **demonstrate** the effectiveness of opt-out mechanisms and technical measures against current agentic scraping capabilities, not merely document their existence. We submit that data protection by design in Article 25(1) GDPR and the regular testing of effectiveness required by Article 32(1)(d) GDPR support this expectation, and we ask the Board to clarify it in the guidelines.
2. Require **periodic reassessment** of scraping mitigations as capabilities evolve. A control that worked in 2024 may be trivially defeated in 2026. The draft already imposes this logic for special-category measures at paragraphs 75 to 77. Extend it to all scraping mitigations.

3. Impose proportionate **due-diligence expectations on acquirers of scraped datasets** on the provenance of the original collection, grounded in the acquirer's existing transparency obligations for data obtained from third parties, including Article 14(2)(f) GDPR.

This is a constructive submission. The guidelines are the right instrument.

Comment 1 (paragraphs 37, 64, 66(f)): documented respect for opt-out signals is not demonstrated effectiveness

The draft correctly identifies website operators' technical measures, robots.txt, ai.txt, CAPTCHA, authentication gates, as signals a compliant controller must heed. Paragraph 37 lists them among pre-collection minimisation measures. Paragraph 64 treats them as shaping data subjects' reasonable expectations, and Example 5 states plainly that where a platform prohibits scraping through robots.txt and CAPTCHA, data subjects cannot reasonably expect their data to be scraped. Paragraph 66(f) adds a discretionary pre-collection opt-out list.

None of these provisions asks whether the measures work.

The draft itself documents the capability class that defeats them. Paragraph 9 describes scrapers that render pages internally, "mimicking a human users' interaction with the page." Paragraph 66(c) notes that a scraping service can receive more data than a regular user, for example through a technical API. An agentic scraping pipeline that rotates identifiers, renders dynamic content, resolves CAPTCHAs, or routes through residential proxies and contracted intermediaries can defeat every signal the draft names, while the controller's legitimate interest assessment records full compliance with paragraph 37. This is the pattern our dataset documents across the adversarial and security incidents it tracks: documented controls defeated by the capabilities they were documented against.

We propose a new paragraph 37bis, in the following terms:

> "Where a controller relies on website operators' technical measures opposing scraping (paragraph 37), on opt-out mechanisms such as the prior right to object (paragraph 66(f)), or on crawler identification (paragraph 30), as part of its lawful-basis assessment or its mitigation package, the

> controller should demonstrate that these measures are effective in practice against current scraping capabilities, including agentic scraping that rotates user agents, IP addresses or other identifiers, renders dynamic content, resolves CAPTCHAs, or routes requests through intermediaries or contracted parties. This follows from the accountability principle in Article 5(2) GDPR and the controller's duty in Article 24(1) GDPR to be able to demonstrate compliance, and from data protection by design in Article 25(1) GDPR: a measure defeated by current capabilities does not implement the principles it purports to serve. Article 32(1)(d) GDPR, which requires regular testing of the effectiveness of technical and organisational measures, illustrates that effectiveness testing is already a GDPR-native expectation. A documented policy of respecting such signals, without evidence of their practical effectiveness, should not be treated as sufficient for the lawful-basis assessment or the mitigation evaluation."
>
> "Controllers should be able to demonstrate effectiveness, for example through adversarial or red-team assessment proportionate to the scale and risk of the processing. Such testing is particularly relevant for large-scale scraping operations and for scraping that may capture sensitive categories of personal data, and should be calibrated to the controller's resources. The results should be documented."

A related correction belongs in paragraph 64. Circumvention of a website's technical measures, whether directly, through contracted parties, or through intermediaries, indicates that data subjects could not reasonably expect the processing, and should weigh heavily against the controller in the balancing test. Example 5 should state this expressly: a scraper that defeats the robots.txt and CAPTCHA it claims to honor has not met the reasonable-expectations standard.

Comment 2 (paragraphs 61, 75-77): extend the reassessment obligation to all scraping mitigations

Paragraphs 75 to 77 already contain the right principle, in the special-categories context: implementation of measures should be "accompanied by regular verification and assessment" of their effectiveness, controllers should "constantly monitor technological developments," and ineffective measures should be replaced. We support this unreservedly. We ask that its logic be generalised.

Scraping capabilities evolve faster than any single assessment cycle. The mitigations a controller documents in a 2024 legitimate interest assessment, static crawler identification, honoring robots.txt, syntax-based filtering, may be trivially defeated by 2026 agentic capabilities. Paragraph 61 instructs that the likelihood of further consequences be assessed "taking into consideration the technical and organizational measures in place." We submit that a measure defeated by current capabilities should not be treated as "in place" for the purposes of paragraph 61. It is a historical artifact.

We propose the following addition after paragraph 66, generalising paragraphs 75 to 77:

> "The verification and reassessment obligation in paragraphs 75 to 77 applies to all technical and organisational measures relied upon under sections 3.2.2 and 3.2.3, not only to measures addressing special categories of personal data. Controllers should reassess the effectiveness of their scraping mitigations at regular intervals, and whenever there is a material change in scraping technology or in the controller's own collection methods, and should document the outcome. Where reassessment shows that a measure is no longer effective against current capabilities, the controller should adopt additional or different measures or revisit the lawful-basis assessment."

Without this, the guidelines will age the way static blocklists age: correctly, and irrelevantly.

Comment 3 (paragraphs 3, 4, 19, 32): deepen the acquirer's provenance obligation

Paragraph 3 brings acquirers into scope: one of the two typical situations the guidelines address is "an organisation obtains and reuses a dataset that has already been scraped by another organisation (data broker)." Paragraph 4 then declines to 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." That is a scope choice, and paragraph 5 confirms it is not an exemption: "This does not mean that the excluded situations are not subject to data protection legislations." A broker that scrapes personal data remains a controller of that

collection, answerable under the GDPR. Paragraph 19 separately allocates responsibility to the acquirer for its own processing: "The scraper is not, in principle, responsible for the reuse of the data."

The gap that remains is narrower than a vacuum but real. Nothing in the draft requires the acquirer to establish the provenance of the dataset it buys. Paragraph 32 merely "recommends" that acquirers link to the original controller's website and explain the conditions of collection. An acquirer can therefore train on a dataset whose original collection disregarded every safeguard in the draft, without any obligation to look back at the circumstances of that collection. The formal allocation of responsibility is clean. The original collection falls outside the guidelines' coordinated safeguards.

We do not ask the Board to bring pure data brokers into scope against paragraph 4's deliberate choice. We ask that the acquirer's side of the transaction carry a real obligation, and we ground that ask in obligations the acquirer already has. The acquirer's own lawful-basis assessment under Article 6(1)(f) GDPR, together with the accountability duty in Article 5(2) GDPR, requires it to understand the circumstances of the collection it builds on: provenance bears directly on reasonable expectations and the balancing test. Transparency reinforces the point: a controller that obtains personal data from a third party must inform data subjects of the source of the data, including whether it came from publicly accessible sources (Article 14(2)(f) GDPR), and the draft itself requires a "precise indication of the source(s)" of the data (paragraph 30). A controller that cannot establish how its training dataset was collected is poorly placed to meet those duties. In the alternative, we present the proposal as a policy recommendation for the final guidelines text.

> "An organisation obtaining an already-scraped dataset for AI development should conduct proportionate due diligence on the dataset's provenance, including whether the original collection respected technical opt-out signals and the data minimisation expectations in paragraph 37, and should document the result. Where provenance cannot be established, or where it shows systematic disregard of opt-out signals, the acquirer should treat the dataset as high-risk: it should re-apply exclusion criteria and opt-out lists before training, and should consider refraining from use where its own lawful-basis assessment cannot be sustained. The principle in paragraph 19

> that the scraper is not, in principle, responsible for re-use is without prejudice to the acquirer's own obligations as controller of its distinct processing."

Data laundering is the predictable response to any regime that prices original collection strictly and acquisition cheaply. The final guidelines should price them the same.

Appendix: the evidence base

The Synthetic Outlaw Observatory scores AI governance incidents on a 10-point severity scale reflecting the gravity of the governance failure, with evidence tiers of verified, probable, candidate, developing, and averted. Current totals: 205 scored incidents (15 verified, 108 probable, 54 candidate, 26 developing, 2 averted); 84 scored 8 or higher; 29 scored 9 or higher; 2 scored 10; 168 logged between June and September 2026. The scoring reflects the Foundation's own methodology. It is published but has not been peer-reviewed or independently validated.

Three incident classes in the dataset bear directly on this submission, cited here in general terms:

- **Agentic systems defeating technical controls at scale.** The dataset includes AI-agent-driven credential theft reported at a scale of 600,000 cards (Gambit Security, September 2026; see e.g. https://gambit.security/blog-posts/autonomous-ai-agents-online-retailers-25-a-company/). The mechanism matters more than the target: autonomous agents defeating authentication, rate limiting, and identity controls that their operators documented as effective.
- **AI compressing the cost of defeating defensive measures.** The dataset includes AI-accelerated attack campaigns in which defensive controls designed against human-speed adversaries failed against machine-speed ones. Opt-out signals designed against conventional crawlers face the same asymmetry against agentic scrapers.
- **Frontier agentic planning of control circumvention.** The dataset includes AI-assisted attack planning scored 10 out of 10: systems reasoning explicitly about how to route around safeguards. A scraper that plans its route around robots.txt is not a malfunction. It is the capability working as

designed.

We state plainly where our evidence is thin. The dataset documents the capability class, agentic AI defeating technical controls, rather than a verified incident of an AI training scraper defeating robots.txt or CAPTCHA specifically. Our application of that capability class to scraping opt-outs is extrapolation from demonstrated capabilities, not a scored scraping incident, and the Board should weigh it accordingly. The dataset also skews toward adversarial and security events rather than GDPR compliance cases. What the evidence does establish, at 205 scored incidents and rising, is that documented controls have repeatedly failed against the capabilities they were documented against in the adversarial and security contexts this dataset tracks. That is the gap these comments address, and the fixes we propose are proportionate to it.

Respectfully submitted,

Jonathan Gropper, JD

Fulbright Specialist in AI governance

Author, *The Synthetic Outlaw: How AI Defeats the Rules Without Breaking Them*

On behalf of The Synthetic Outlaw Observatory Foundation (501(c)(3))