Comment on EDPB Guidelines 03/2026 on Web Scraping in the Context of Generative AI

Submitted by: Sarah McKenna, in an individual capacity

Consultation closes: October 30, 2026, 23:59 CET

1. Who I am and why I am commenting

I have spent most of the last decade running and building web data collection systems. I spent nearly nine years at Sequentum, a web data platform serving institutional finance, including more than seven years as CEO and a board member, and before that I worked in security, QA and technical operations at Microsoft and elsewhere. I was the sole technical advisor to the SIIA/FISD Alternative Data Council's *Web Data Collection Considerations* series, which covers robots.txt, CAPTCHA, privacy, terms of service and related topics for the investment industry. I am a co-author of the individual IETF Internet-Draft *Best Practices for Responsible Web Data Collection* (draft-farzdusa-webbot-datacollection), and I have presented on data collection use cases at IETF 124 and at an IETF Web Bot Auth interim session. I serve on the steering committee of the Alliance for Responsible Data Collection (ARDC), which is submitting its own comments. The views here are my own.

I support the Board's goal. Collection of personal data for AI training should be minimized, filtered, transparent and accountable. My concern is narrower and, I think, fixable: the draft treats a website operator's machine signals (robots.txt, ai.txt and CAPTCHA) as if they were evidence of what the people whose data appears on that site expect. From where I sit, operating these systems every day, they are not, and treating them that way will weaken privacy protection rather than strengthen it.

2. The provisions at issue, and why they are weak

The four passages below carry the same idea. Each is quoted as written in the draft, followed by my explanation. The evidence behind each point is in section 5.

Paragraph 11 (first step of the scraping process):

> "This step also implies implementing measures to exclude certain sources from the collection (e.g. websites which oppose web scraping with measures such as robots.txt or CAPTCHA)."

Why this is weak: It defines "opposing scraping" by the tool a site happens to deploy, not by anything a data subject has said or done. robots.txt is, by its own IETF standard, a request and not an access control. CAPTCHA is most often a traffic-management device. Official public registers use it as a speed bump (see the Delaware example in 5.5). Neither tells a controller anything about personal data.

Paragraph 37, sixth bullet (data minimization, before collection):

> "exclude from the collection websites which clearly oppose the scraping of their content, through the use of technical measures, such as the use of robots authentication to view content, robots.txt or ai.txt files, or CAPTCHA, which impose an action that can only be carried out by a human being and aim to prohibit access to pages by robots."

Why this is weak: Data minimization under Article 5(1)(c) asks whether particular personal data is necessary for a purpose. This bullet replaces that question with a different one: whether the site owner dislikes automation. A price page with no personal data would be excluded; a forum full of personal data with no robots.txt would not. The test points in the wrong direction. Its factual premise is also out of date. CAPTCHA does not reliably "impose an action that can only be carried out by a human being": bots now solve it more accurately than people do. I have no objection to excluding content that sits behind a login. That content is genuinely not public, and paragraph 66(c) already covers it.

Paragraph 64, third bullet (reasonable expectations of data subjects):

> "the restrictions imposed by the scraped website, for example through the implementation of technical measures such as the use of robots authentication to view content, robots.txt or ai.txt files, or the implementation of CAPTCHA, which are intended to impose an action that can only be carried out by a human being, and aim to prohibit access to the pages by robots. If data subjects are aware that a website implements these measures and a website is nevertheless scraped, it is less likely that they can expect the processing of their personal data by a scraping entity."

Why this is weak: Reasonable expectations belong to the data subject, but this bullet derives them from the operator's server configuration. Most people have never heard of robots.txt and do not control it. The bullet also conditions its conclusion on awareness ("If data subjects are aware…") without asking whether awareness is realistic. It then rewards a belief that is false: that a text file keeps public posts from being collected.

Example 5:

> "If a person uploads their data to a content sharing platform that prohibits scraping through the use of robots.txt files and the implementation of CAPTCHA, and expressly states on their site that they do not allow the use of their users' data for

the development of AI models, they cannot reasonably expect third parties to scrape that data for that purpose."

Why this is weak: The only element here that speaks to users is the platform's express statement. robots.txt and CAPTCHA add nothing to it. And platforms routinely say one thing in robots.txt while licensing the same user content for AI training under their terms (the Reddit example in 5.3). The example should turn on what users were actually told and what they control.

3. Summary of recommendations

1. Delete the parenthetical in paragraph 11.
2. Remove the sixth bullet of paragraph 37 and the third bullet of paragraph 64, or reword them so that operator anti-automation signals are not treated as data-subject expectations or as a data minimization criterion. Exclusion of content behind a login should stay.
3. Revise Example 5 so that reasonable expectations turn on what the data subject was told and controls, not on the platform's crawler configuration.
4. Strengthen the measures that genuinely track data-subject interests: user-level privacy settings, the opt-out list in paragraph 66(f), PII and special-category filtering, and published source lists.
5. Add the public's right to receive information and the effect on competition to the balancing test.

Proposed replacement text is in section 7.

4. Where I agree with the draft

- The targeted/untargeted distinction in paragraph 8 is right, and it matters. Most commercial collection is narrowly targeted at prices, inventory, filings and similar non-personal data.
- Defining collection criteria in advance, PII filtering, pseudonymization and timestamping (paragraphs 37–38 and 42) are good practice. They mirror what responsible operators already do and what my co-authors and I describe in our Internet-Draft.
- Paragraph 19 correctly says a scraper is not, in principle, responsible for a downstream AI developer's reuse.
- Paragraph 45 correctly says the absence of a robots.txt file is not consent.

5. The evidence: an operator's crawler settings are not a data subject's expectations

5.1 A Disallow is a real "no." The questions are whose, and to what

I do not claim that a robots.txt Disallow means nothing. The absence of a file is silence; a Disallow is a statement. But three questions decide how much weight that statement deserves under the GDPR:

- **Who is speaking?** The website operator, who is often also the rightsholder. It is not the data subject. Operators block automation for commercial, competitive and operational reasons that may have nothing to do with their users' privacy.
- **What does it cover?** A Disallow applies site-wide, to every purpose. It sweeps in prices, filings, product data and other information the GDPR does not govern, and it cannot single out personal data.
- **Is it consistent?** The same operator may license its users' content for AI training under its terms of service while disallowing everyone in robots.txt (5.3), or block every crawler except the dominant search engine (5.4).

When the answers are "the operator," "everything," and "it depends who is asking," the file expresses a business preference. It is not evidence of what individuals expect, and it is not a data minimization criterion. Copyright law can give a rightsholder's Disallow legal effect (5.2); data protection law should look to the data subject.

5.2 robots.txt is, by its own definition, a request rather than a control

The IETF standard that defines robots.txt, RFC 9309, states plainly that its rules "are not a form of access authorization." Courts read it the same way. In December 2025, in *Ziff Davis v. OpenAI* (S.D.N.Y.), Judge Stein held that robots.txt directives do not effectively control access to content any more than a "keep off the grass" sign controls access to a lawn, and that ignoring them is not circumvention. US courts have also repeatedly declined to let platforms convert public visibility into private control. In *X Corp. v. Bright Data* (N.D. Cal. 2024), Judge Alsup relied on the Ninth Circuit's warning in *hiQ Labs v. LinkedIn*, 31 F.4th 1180 (9th Cir. 2022), that giving platforms free rein over who may collect public data risks "information monopolies that would disserve the public interest."

In the EU, robots.txt does have one recognized legal role: as a machine-readable way for **rightsholders** to reserve text-and-data-mining rights under Article 4(3) of the DSM Directive (see *Kneschke v. LAION*, OLG Hamburg, 5 U 104/24, December 10, 2025). That supports my point. The law treats robots.txt as the voice of the rightsholder or operator. Nothing in that framework treats it as the voice of a natural person whose comment, photo or profile happens to sit on the site.

5.3 In practice, robots.txt is coarse, stale and frequently contradictory

I have read thousands of robots.txt files professionally. They are usually written once, copied from a template, and rarely revisited. The research matches my experience:

- MIT's Data Provenance Initiative (*Consent in Crisis*, NeurIPS 2024) found robots.txt files that allow some AI companies while blocking others, name crawlers that do not exist,

and contradict the same site's terms of service. The authors concluded that the protocol forces operators to anticipate every agent in advance, which produces files that don't reflect intended consent.
- The same study found that where any AI crawler is restricted, OpenAI's crawler is disallowed 91.5% of the time, but Google Search only 17.1% of the time.
- A file is site-wide and agent-based. It cannot say "don't collect personal data, but prices are fine," which is the distinction the GDPR actually cares about. ai.txt, which the draft names alongside robots.txt, is weaker still: it is not a recognized standard and is not broadly adopted, or even known, among website operators.

Reddit shows how far robots.txt and a platform's actual data practices can diverge. Its robots.txt disallows all automated agents ("User-agent: * Disallow: /"). Yet its User Agreement, Section 5 ("Your Content"), version effective July 1, 2026, as viewed October 19, 2026, states: "For example, this license includes the right to use Your Content to train AI and machine learning models, as further described in our Public Content Policy." Reddit has licensed user content to Google and OpenAI for AI training. A Reddit user who relied on the robots.txt would be wrong about what happens to their data.

5.4 The rule would protect incumbents, not people

This is my most serious concern. A common robots.txt pattern is "disallow everyone except Googlebot." An analysis of robots.txt files across Cloudflare's network in Q1 2026 found Googlebot named in allow rules more than three times as often as in disallow rules. Meanwhile, Google's AI Overviews are generated from content collected by the same Googlebot crawl. The UK Competition and Markets Authority has found that, given Google's strategic market status in search, publishers have had no realistic choice but to allow that crawl.

Put those facts together with paragraphs 37 and 64 as drafted. The same personal data, on the same page, would remain available for generative AI to the dominant search incumbent and to anyone who can afford a licensing deal. It would be off-limits to an EU start-up, a university lab, or a small company in an emerging market. The data subject's privacy is identical in both cases. The only thing the rule changes is who gets to compete. I do not believe that is the Board's intent, but it would be the effect.

5.5 CAPTCHA is a traffic and anti-abuse tool, and it no longer tells humans from machines

CAPTCHA exists to manage fraud, spam and server load. It says nothing about the privacy expectations of the people whose information sits behind it.

One example I know well is the Delaware Division of Corporations' entity search. Delaware is where most US companies are incorporated. The State provides basic entity information free of charge to the public: entity name, file number, formation date and registered agent. This is a public register of exactly the kind paragraph 7 of the draft lists. Yet the search sits behind a CAPTCHA, and the page states that the Division prohibits data mining. The CAPTCHA is not

there to protect anyone's privacy. It works as a speed bump, slowing traffic so the system is not overloaded. Under paragraph 37 as drafted, a controller would have to treat an official public register, whose purpose is public access, as a site that "clearly opposes" collection.

CAPTCHA also no longer works as a human test:

- A UC Irvine study presented at USENIX Security 2023 had 1,400 participants solve 14,000 CAPTCHAs drawn from popular websites. Bots solved them with up to 99.8% accuracy versus 50–84% for humans, and in under a second versus up to 15 seconds.
- The W3C has documented for years that CAPTCHA excludes many people with disabilities and many non-English speakers.
- TollBit reported in 2025 that consumer AI browsers and agent tooling now load pages and ads and solve CAPTCHAs while appearing in site logs as ordinary Chrome users.

A test that machines pass more reliably than people, and that disabled users often cannot pass at all, is not a sound basis for inferring what a data subject expects.

5.6 The web is already majority-automated, and much of that automation acts for people

- Imperva's 2025 Bad Bot Report found automated traffic reached 51% of all web traffic in 2024, the first time in a decade it exceeded human traffic.
- HUMAN Security's 2026 benchmark found AI agent and agentic-browser traffic grew 7,851% year over year in 2025. DataDome reported agentic browser requests on its network jumped 4.7x in September 2026 alone.

Much of this is a person asking an assistant to read, compare or buy something on their behalf. Guidance that equates "opposes automation" with "data subjects expect no processing" will age badly as more ordinary browsing is delegated to agents. It also keys GDPR outcomes to a signal that even large AI operators interpret differently. TollBit's H1 2026 data found roughly 15% of AI scrapes across its European and North American publisher cohorts bypassed an explicit disallow, and OpenAI states that robots.txt may not apply to user-initiated fetches.

5.7 Telling people a text file protects them is a disservice to them

Paragraph 64 says that if data subjects are aware a site uses these measures, they are less likely to expect scraping. My worry is what that teaches people. Content posted publicly is indexed by search engines, preserved by archives, licensed by platforms, copied by humans, and collected by automated tools that do not all comply. The figures in 5.6 show non-compliance is measurable and persistent.

A data protection authority should tell people the truth about how the web works: if you post something publicly, assume it can be copied, and look to the law and the controller's obligations, rather than a platform's crawler file, for protection. Encouraging a belief that robots.txt shields public posts creates a false sense of security, which is the opposite of transparency. The real

protections are the ones the draft already describes well: minimization, filtering, special-category safeguards, transparency and enforceable rights.

5.8 The balancing test is missing two interests

The draft balances the controller against the data subject. Articles 11, 13 and 16 of the Charter also protect the freedom to receive and impart information, the freedom of the arts and sciences, and the freedom to conduct a business. Price comparison, market transparency, academic research, journalism, archiving and new market entrants all depend on collecting public information. Guidance that lets private operators decide, through a configuration file, which lawful collectors may access public information hands an information-access gatekeeping role to the largest platforms. I ask the Board to name these interests in the balancing test and to weigh the competitive effects described in 5.4.

5.9 Much legitimate collection cannot announce itself

Robots.txt only works for collectors willing to name themselves: an operator can allow or block a crawler only by its declared name. Many of the most socially valuable collectors cannot do that. If the subject of the study knows who is watching, it changes what it shows. That is the point of studying it unannounced. Europe learned this lesson with Dieselgate: a system that can tell it is being tested behaves differently. The draft itself recognizes, in paragraph 66(c), that sites can return different outputs depending on the user agent. In my own work I have seen sites send false data, such as altered prices or empty pages, to collectors they recognize.

Legitimate collectors that need to stay unannounced include:

- **NGOs tracking hate speech** and how it targets particular regions and groups, and the detection models they train on what they collect.
- **Investigative journalists and fact-checkers.**
- **Businesses and price-comparison services tracking competitors' prices**, which are often personalized or altered for known bots.
- **Consumer organizations** testing what an anonymous person is actually shown, for example personalized or discriminatory pricing.
- **Government watchdogs and regulators.** The EU does this itself. The Commission coordinates CPC sweeps, including an AI-assisted review of 16,000 products on 176 websites, and provides market surveillance authorities with a web crawler to find unsafe products online. The European Parliament's 2025 resolution on product safety notes that authorities sample products online under a covert identity, as "mystery shoppers."
- **Political campaign and election-integrity monitors** tracking disinformation, political advertising and coordinated inauthentic behavior.
- **NGOs and technology vendors supporting police counter-trafficking and child-protection work.**
- **Academic and algorithmic-accountability researchers** auditing platforms. The Digital Services Act (Article 40(12)) already recognizes researchers' need to collect publicly accessible platform data.

- **Investors tracking market trends** through alternative data.
- **Ad verification and ad-fraud detection**, confirming that ads appear where they were paid for and do not carry malware.
- **Brand protection and counterfeit detection** on online marketplaces.
- **Cybersecurity researchers** finding phishing and scam sites, which routinely show harmless content to scanners they recognize.
- **Sanctions, anti-money-laundering and adverse-media screening** by compliance teams.
- **Web archivists** preserving the public record.

None of these collectors will ever be named in a robots.txt allow list. Under paragraph 37 as drafted, each would be told to stay away from any site that would rather not be observed. Those are often the sites that most need observing.

The answer is not to make every collector identify itself to its target. It is **anonymous attestation of good behavior**. A collector should be able to prove that it belongs to a class of audited, well-behaved bots without revealing who it is or whom it works for. Well-behaved means respecting rate limits, never bypassing logins, filtering personal data, and honoring data-subject signals. Accredited auditors would certify the operators, and privacy-preserving cryptography, of the kind already used in Privacy Pass, would carry the proof. This is under active discussion at the IETF, including at the Web Bot Auth sessions where I presented these use cases. Anonymity toward the target does not mean anonymity toward the law: the auditor, and a supervisory authority with a lawful request, would know who the operator is. Article 14 transparency can still be met through the controller's public notice and source list.

I ask the Board to recognize audited, anonymous attestation as a mitigating measure under paragraph 66. I also ask it to state that a collector's choice not to identify itself to a website is not, by itself, evidence of bad faith.

6. What would actually track data-subject interests

I want to offer alternatives, not only objections.

- **Honor signals the data subject controls.** Many platforms let users mark a profile or post as private, not indexable, or not for AI use. That is a genuine expression of the individual's expectation, and the Board could say clearly that controllers should respect it.
- **Honor specific, user-facing promises.** If a platform's terms tell users their public personal data will not be scraped, sold or licensed, that is relevant to reasonable expectations. A blanket crawler rule is not.
- **Build out paragraph 66(f).** A workable, pre-collection opt-out list lets individuals object directly, without depending on a platform's choices.

- **Keep the strong technical measures.** PII filtering, special-category exclusion, exclusion of sites used mainly by minors, pseudonymization, deduplication and published source lists do far more for individuals than site-level blocking.
- **Look to purpose-specific standards rather than freezing a 1994 convention into guidance.** The IETF is working on purpose-specific preference vocabularies (AIPREF) and on ways for well-behaved automated clients to identify themselves (Web Bot Auth). The Board's guidance should stay technology-neutral so it can recognize better signals as they mature, and should prefer signals that distinguish personal data and purpose over blanket agent blocks.

7. Proposed text changes

Paragraph 11. Delete: "(e.g. websites which oppose web scraping with measures such as robots.txt or CAPTCHA)."

Paragraph 37, sixth bullet. Replace with:

> *exclude from the collection personal data that the data subject has designated as private or not for reuse through settings made available to them by the website, and content available only after logging in; and consider excluding websites whose terms clearly inform users that their publicly posted personal data will not be scraped, sold or licensed. A website operator's general preferences regarding automated access, including those expressed through robots.txt, ai.txt or CAPTCHA, do not by themselves indicate the expectations of data subjects or determine whether particular personal data may lawfully be processed.*

Paragraph 64, third bullet. Replace with:

> *the information and controls made available to data subjects by the source website, for example whether users were told their public posts would not be scraped, sold or licensed, and whether they used available privacy or no-reuse settings. Operator-level anti-automation measures such as robots.txt, ai.txt or CAPTCHA are not in themselves indicative of data subjects' expectations.*

Example 5. Replace with:

> *A person posts content on a non-public portion of a platform, such as a group or page visible only to people they chose, and the platform's terms state that such content will not be scraped, sold or licensed without the user's consent. Or a person posts publicly but enables a "do not use my content for AI" setting that the platform offers and honors. In either case, the data subject can reasonably expect that third parties will not process that content for AI development.*

8. Closing

The draft gets most things right, and the Board has an opportunity to make Europe's guidance the clearest in the world on responsible collection. My request is simple: keep the focus on personal data and on the people it describes. An operator's crawler file and a traffic-control puzzle are tools for managing servers and businesses. Treating them as proxies for individual expectations would leave individuals no better protected, mislead them about how their public data travels, and hand a competitive advantage to the companies that already hold the most data.

I would welcome the chance to discuss any of this with the Board's experts.

Sarah McKenna, Brooklyn, NY / https://www.linkedin.com/in/sarahransommckenna/

Appendix: sources

Under comment

- EDPB, Guidelines 03/2026 on web scraping in the context of generative AI, v1.0 (adopted July 7, 2026); consultation page

Standards

- IETF, RFC 9309, Robots Exclusion Protocol (September 2022)
- W3C, Inaccessibility of CAPTCHA

Court decisions

- *Ziff Davis v. OpenAI*, S.D.N.Y., December 2025, order on robots.txt and DMCA §1201 (PDF)
- *hiQ Labs, Inc. v. LinkedIn Corp.*, 31 F.4th 1180 (9th Cir. 2022) (PDF)
- *X Corp. v. Bright Data Ltd.*, No. C 23-03698 WHA, N.D. Cal., May 9, 2024 (PDF)
- *Kneschke v. LAION e.V.*, OLG Hamburg, 5 U 104/24, December 10, 2025, case summary

Platform and government sources

- Reddit, User Agreement, Section 5 (version effective July 1, 2026, as viewed October 19, 2026); Reddit's own filing reproducing Section 5, AB 587 Terms of Service Report H2 2025
- Reddit, robots.txt (as viewed October 19, 2026)
- Delaware Division of Corporations, General Information Name Search
- UK Competition and Markets Authority, proposed conduct requirements for Google search

Research and industry data

- Longpre et al., *Consent in Crisis: The Rapid Decline of the AI Data Commons*, NeurIPS 2024
- Searles et al., *An Empirical Study & Evaluation of Modern CAPTCHAs*, USENIX Security 2023
- Imperva (Thales), 2025 Bad Bot Report
- HUMAN Security, 2026 State of AI Traffic & Cyberthreat Benchmark Report
- DataDome, AI Traffic Report Q3 2026 (October 15, 2026)
- TollBit, State of the Bots Q2 2025; H1 2026 findings as summarized here
- Analysis of robots.txt allow/disallow rules across Cloudflare's network, Q1 2026

IETF materials and EU enforcement practice

- Sarah McKenna, *Mission-Critical Web Data & the Bot-Auth Threat*, slides, IETF 124 Web Bot Auth session, Montréal, November 4, 2025; session minutes
- European Commission, CPC consumer protection sweeps
- European Parliament, resolution on product safety and market surveillance (2025)
- Industry Europe, Commission launches "web crawler" for unsafe products (Safety Gate 2021 report)