bright data

Bright Data welcomes the EDPB's effort to clarify how the GDPR applies to web scraping in the context of generative AI. Bright Data provides public web data collection infrastructure and services to a wide range of organizations and therefore has a direct interest in a clear and workable framework. The Guidelines rightly recognize that web scraping can be a lawful and valuable data collection activity; that Legitimate Interest is often the most relevant legal basis for private entities to rely upon; and that individualized notice required by Article 14 may prove technically impossible or requires unreasonable, disproportionate effort.

Nevertheless, Bright Data proposes targeted revisions in two key areas so that the Guidelines more accurately reflect the state of the industry and the importance of preserving access to public web data.

Public vs. Nonpublic. The Guidelines correctly distinguish between data subjects' reasonable expectations of privacy regarding publicly available data and their expectations regarding non-public or access-restricted data (paragraph 64) and already recognize limiting collection to freely accessible data as a mitigating measure (paragraph 66(c)) and, in Example 7, that the balancing test may generally be met where only freely and publicly accessible sources are used. We welcome this distinction. However, we believe the distinction is fundamental and should be given greater and more consistent weight throughout the Guidelines, including in the necessity and proportionality analysis. Publicly available web data should be understood, at minimum, as data that is accessible to any internet user without payment, log-in credentials, membership approval, account creation, contractual onboarding, or circumvention of technical access controls. This category is materially different from data obtained from closed groups, private accounts, paywalled services, credentialed platforms, or restricted APIs. We note that paragraph 64 frames its standard around data made public "by the data subject themselves". Much public web content concerning individuals is legitimately published by third parties, such as news outlets, public registers and business directories. We invite the EDPB to clarify that where data is openly accessible on the web and was lawfully published, the public character of the source remains a relevant factor favoring the controller, even where the publisher is not the data subject, while accepting that the specific cases identified in paragraph 66(c), where data subjects cannot reasonably be aware of the accessibility, may warrant additional care.

This distinction matters because the privacy risks, reasonable expectations, and proportionality analysis are different where information has been made available on the open web. In many cases, individuals, businesses, public bodies, publishers, and other organizations intentionally make limited information publicly accessible so that it can be discovered, read, indexed, referenced, compared, analyzed, or otherwise used by third parties. The open web depends on that possibility. Search engines, archiving services, cybersecurity tools, price-comparison services, market-intelligence tools, academic research, fraud detection, brand-protection services, AI safety tools, and many other socially and economically valuable services rely on the ability to access and process publicly available web data.

As mentioned in the Guidelines, the public availability of personal data does *not* mean that the GDPR is inapplicable, nor does it eliminate the need for safeguards. Public availability should,

however, be expressly recognized as a mitigating factor not only in the balancing test, under Article 6(1)(f), but also when assessing the necessity of the processing (paragraphs 51-52), the reasonable expectations of data subjects and the appropriate level of transparency and data-minimization measures.

Limiting collection to freely accessible public web data should weigh in favor of the controller in the legitimate-interest balancing test, provided that the controller implements appropriate safeguards. Consistent with the mitigating measures already identified in paragraph 66, such safeguards may include avoiding paywalled or logged-in environments, excluding sources that are likely to contain sensitive or children's data, applying reasonable filtering and deduplication measures, maintaining source provenance at an appropriate level, publishing clear public transparency notices, providing practical mechanisms for data-subject requests, and deleting, suppressing, or anonymizing data where appropriate. We suggest that Example 7 be generalized into express guidance that this combination of measures will generally support reliance on Article 6(1)(f).

The Guidelines rightly distinguish between targeted and untargeted scraping (paragraph 8) and link impact to volume and indiscriminate collection (paragraph 57).. We agree that untargeted collection warrants greater care. However, the type of source should carry as much weight as scale: a company that collects only information available to the general public under defined collection criteria is not comparable to an actor that collects personal data that is not public. The Guidelines should therefore emphasize that the use of open public web sources is a significant factor in the assessment of reasonable expectations and proportionality.

Technical Protocols and Anti-Bot Technology. The Guidelines identify robots.txt, ai.txt, CAPTCHAs, and similar measures as indicators that a website "clearly opposes" scraping (paragraphs 11 and 37) and as restrictions relevant to reasonable expectations (paragraph 64). Bright Data accepts that authentication requirements such as log-in walls, are a strong indicator that content is not intended for open access, and that such sources should generally be excluded. However, the treatment of robots.txt, ai.txt and CAPTCHAs is oversimplified: a generic measure of this kind does not, on its own, amount to *clear* opposition to the processing of personal data and treating it as such risks undermining legitimate uses of public data.

CAPTCHAs, for example, are primarily access-management and abuse-prevention mechanisms. They are commonly deployed to prevent spam, credential stuffing, denial-of-service attacks, fake-account creation, automated form submissions, excessive traffic, scraping at abusive rates, or other forms of malicious automation. Their presence does not necessarily indicate that the website operator objects to all automated access and less so that the relevant data subjects object to processing under the GDPR. Moreover, CAPTCHAs do not distinguish between harmful automation and legitimate automated collection, such as crawling for search, academic research, cybersecurity, fraud detection, compliance monitoring, accessibility, market intelligence, or public-interest analysis. A CAPTCHA may therefore signal a concern about the manner, volume, or security implications of access, rather than a privacy-based objection to the processing of personal data.

Robots.txt serves an even narrower function than CAPTCHA. It was originally developed as a voluntary crawl-management protocol for web crawlers, particularly search engines, and historically addressed operational concerns such as server load, bandwidth use, indexing preferences, and crawler etiquette. While it may be relevant in other legal contexts, it is not a GDPR

rights-management tool and was not designed to capture nuanced data-protection preferences. A robots.txt instruction may reflect a site operator's technical or commercial preference about crawling or indexing, but it does not necessarily express a data subject's reasonable expectation, objection, or privacy interest. In many cases, the website operator configuring robots.txt is not the data subject whose information appears on the page. Paragraph 64 itself conditions the relevance of these measures on data subjects being aware that a website implements them; in practice, very few individuals know whether, or how, the websites on which their information appears use robots.txt, ai.txt or CAPTCHA.

Ai.txt is even less settled. It is an emerging, non-universal convention aimed at AI-related uses, but it lacks broad legal recognition, standardization, consistency, and adoption. To the extent ai.txt is used, it may reflect a publisher's copyright or licensing position, business strategy, bargaining objective, objection to model training, desire to monetize content, concern about attribution, or broader policy preference regarding AI. We recognize that Example 7 refers to text and data mining opt-outs under Directive 2019/790; those opt-outs are properly addressed under copyright law. Such concerns may be important in other legal or commercial contexts, but they should not automatically be translated into GDPR conclusions about reasonable expectations, legitimate interest, or lawfulness of processing personal data.

Converting website-owner preferences into GDPR rules would allow platforms and publishers to transform technical or commercial signals into de facto veto rights over otherwise lawful access to public web data. That could undermine legitimate uses of public information, including search, research, cybersecurity, fraud detection, competition monitoring, price comparison, public-interest investigations, AI safety, and other socially valuable activities. It could also entrench the position of large platforms that can selectively permit preferred crawlers while excluding competitors or smaller innovators.

Proposed Approach. Rather than treating all technical protocols alike, Bright Data proposes that the Guidelines adopt a graduated approach: (i) content protected by authentication or paywall requirements should generally be excluded; (ii) a specific instruction directed at AI training or at the relevant crawler, or an express statement by the website that its users' data may not be used for AI development (as in Example 5), should be a relevant factor, but should not automatically be translated into GDPR conclusions about reasonable expectations, legitimate interest, or lawfulness of processing personal data; and similarly (iii) generic robots.txt directives or abuse-prevention CAPTCHAs should be one contextual factor among others. We suggest amending paragraphs 11, 37 and 64, and the executive summary, accordingly.

Conclusion. Bright Data supports the EDPB's objective of ensuring that web scraping for generative AI respects data subjects' rights. We respectfully ask that the final Guidelines (i) give consistent weight to the public availability of sources throughout the Article 6(1)(f) analysis, and (ii) treat robots.txt, ai.txt and CAPTCHAs as contextual factors rather than determinative indicators of opposition to processing. We would welcome the opportunity to engage further with the EDPB on these points.