Arcom
The Regulatory Authority for Audiovisual and Digital Communication

July 2023

# The essentials

## Review and outlook for the fight against the spread of online hateful content: *making online platforms more accountable, with a view to the implementation of the Digital Services Act*

**The French law of 24 August 2021 reinforcing respect for the principles of the Republic** strengthened the legal framework governing the responsibility of the main online platform providers active in France for combating the spread of hateful content on their services.

The aim of this national framework was to anticipate the **European regulation on digital services** ("*Digital services act*", or DSA), adopted on 19 October 2022, from which it takes its inspiration and which will replace it in early 2024 in keeping with the principle of maximum harmonisation. For the largestonline platforms, the DSA will come into force on 25 August 2023 under the super- vision of the European Commission.

Arcom's review of the situation provides an opportunity to test the framework. **Based on observations and statements made by providers in response to an *ad hoc* questionnaire**, it reports on improvements to the tools and procedures put in place to combat misuse of the main online platforms.

## A transparency drive still to be consolidated

**On the whole[1], the providers questioned demonstrated greater transparency.** The statements provide high-quality, previously unpublished data, particularly on the number and origin of reports of hateful content, the human resources used for moderation and their collaboration with the public authorities and trusted reporters.

In particular, some companies, such as *Dailymotion* and *Meta*, have developed relevant tools and indicators to **assess the prevalence of** unlawful content on their services. Their efforts contrast with the shortcomings of services such as *Twitter* and *Bing* (*Microsoft*).

Their transparent approach reflects a **genuine willingness to prepare for compliance with the DSA,** under which VLOPSEs[2] will have to demonstrate enhanced transparency and be required to assess the systemic risks associated with their platforms and undertake steps to mitigate them.

**The problems encountered by some platforms[3]** in gathering and transmitting detailed data within the allotted time attest to the need to implement analysis and reporting resources for this purpose. There is also a need for efforts to improve transparency with regard to the number and working language of the moderators responsible for fighting the spread of online hateful content, as this general-interest information is still too often subject to requests for confidentiality, to the detriment of public information.

1 *Pinterest* is the exception.
2 Very large online platforms and search engines exceeding 45 million monthly active addressees in the European Union, and designated as such by the European Commission.
3 For example, *Google and TikTok*.

- Table showing human resources of content moderation

| *Table showing* ***human resources for moderating*** *content*[4] | Number of people assigned to handling online hate reports and appeals from users of the service's French version | Location of people assigned to handling online hate reports and appeals from users of the service's French version | Working language(s) of the people assigned to handling online hate reports and appeals from users of the service's French version |
|---|---|---|---|
| *Bing* | *"Microsoft has a variety of content moderation frameworks with hundreds of full-time employees, vendor employees and occasional employees."* | *"Located in numerous locations worldwide."* | *"Bing ensures that the content moderation teams responsible for reviewing complaints in French have adequate staff to deal with the French language."* |
| *Dailymotion* | Average of around thirty people | European Union (France and Romania) | *"A wide variety of languages"* |
| *Google Search* | Data not provided | Data not provided | Data not provided |
| *LinkedIn* | *"Hundreds of content reviewers [...] including around 29 French language specialists"* (e.g.: by the end of 2022, 278 full-time employees based in Europe, the Middle East and Africa and contractors) | *"Located worldwide"* | *"With local language expertise"* |
| *Meta* (*Instagram* and *Facebook*) | **Confidential** | **Confidential** | *"Including French, English, German and Turkish. "* |
| *Pinterest* | **Confidential** | **Confidential** | **Confidential** |
| *Snapchat* | **Confidential** | **Confidential** | **Confidential** |
| *TikTok* | **Confidential** | **Confidential** | **Confidential** |
| *Twitter* | *"149 people, including French-speakers"* | *European Union ("and based in Ireland and Portugal")* | *"including French-speakers"* |
| *Wikipedia* | *"A geographically distributed team of around 45 to 50 people was available [...] most cases of this kind are handled by two to six people."* | *"Geographically distributed team"* | *"Inherently multilingual* |
| *Yahoo portal* | Data not provided | Data not provided | *"including French-speaking lawyers if necessary"* |
| *YouTube* | Data not provided | Data not provided | Data not provided |

## More must be done to improve the clarity and accessibility of general terms and conditions (GTCs)

The degree of accessibility of the GTCs, from the main page of the services, appears satisfactory via a web browser on a computer (three clicks maximum). However, access is often slightly more complex on mobile phones, via a web browser, with a maximum of four clicks required. Some services are even less accessible via the application, even though this is the main means of access for the vast majority of them (five clicks maximum).

Given the nature of this type of content, offering clear GTCs is both a challenge and a necessity. The GTCs of the services observed are all available in French, with the exception of *LinkedIn,* where some sections are in English only. Nevertheless, for several of them, the layout of the GTCs could be improved to make them easier to understand. A good practice that some providers have implemented in anticipation of the DSA is to provide summaries of each section and to highlight them visually within the GTCs.

Similarly, the GTCs or community rules of most of the services observed mention the possibility of accessing an internal system for handling complaints about a moderation decision. However, three platforms[5] stand out for not explicitly mentioning this option in their GTCs.

[4] Information valid as at December 2022.
[5] *Pinterest*, *Snapchat* and *Bing*.

# Reporting content and challenging moderation decisions

The Authority notes that, **on the whole, providers have taken steps to encourage the reporting of wrongful content on their platforms** and to involve their users. With the exception of some[6], they also allow users who are not logged into an account to report illicit content. Some services[7] offer an *ad hoc* form for reporting hateful content under article 6-4 of the law of 21 June 2004 on confidence in the digital economy (LCEN)[8].

However, on many services, **access to the reporting system requires clicking on a button whose title is not very explicit** (e.g.: "…").

Arcom therefore urges the platforms to pay particular attention to the accessibility and comprehensibility of these systems.

- Table of reports from service users[9]

| *Table showing reports from service* ***users*** | Declared number of reports of hateful content within the meaning of article 6-4 of the LCEN received in 2022 | Number of hateful content moderation actions in 2022 | Percentage of moderation actions taken following a human decision |
|---|---|---|---|
| *Bing* | 0 | 0 | Data not relevant |
| *Dailymotion* | 387 | 490[10] | 100% |
| *Google Search* | Data not relevant | Data not provided | 100% |
| *LinkedIn* | **Confidential** | 3,064 | **Confidential** |
| *Meta* (*Instagram* and *Facebook*) | 12424 | 4,807[11] | 100% |
| *Pinterest* | **Confidential** | **Confidential** | **Confidential** |
| *Snapchat* | 24,184 | 5,637[12] | 100% |
| *TikTok* | 395,302 | 152,628 | 37.6% of deletions based on a breach of community rules, 100% of geo-blockings and deletions following reports made via the "Report hateful content in France" system |
| *Twitter* | 1,159,206 out of a total of 3,405,570 alerts[13] received | Measures taken 289,169 items (action rate: 24.95%) | 34% of all reports[14] |
| *Wikipedia* | 0 | 0 | *"Not applicable"* |
| *Yahoo portal* | *"0 (N/A)"* | 0 | Data not relevant |
| *YouTube* | Data not relevant | Data not provided | 100% |

[6] *Twitter*, *Pinterest* (except via an ad hoc form) and *LinkedIn* do not allow their offline users to report illegal content.
[7] Those of *Google Search*, *YouTube*, *TikTok*, *Facebook* and *Instagram* are accessible to everyone; the *Pinterest* site is only accessible to logged-in users.
[8] Resulting from the law of 24 August 2021.
[9] Arcom urges these figures to be analysed with caution: they are not in themselves sufficient to assess the efforts made by the platforms, and must be considered in relation to platform's size, traffic and model, the completeness and clarity of its reporting system, etc.
[10] The operator's contribution specifies that several moderation decisions may apply to the same content.
[11] Among these, the operator's contribution states that 19 items of content did not violate its Community Standards, but were blocked in France because of a violation of a provision of article 6-4 of the LCEN.
[12] The operator also states that it has taken action against 4,119 single accounts for breaching its hate speech policies in France.
[13] The operator's contribution specifies that this idea covers user reports and *machine* learning reports.
[14] As a reminder: 3,405,570.

**The option of challenging moderation decisions, which is available to users almost everywhere,** is still used unevenly on different platforms. In addition, while it is difficult to draw general conclusions from this situation, the particularly high rate of reversal of initial moderation decisions for *TikTok* (40.8% of appeals following a report via the system dedicated to violations of the Community Rules) and Dailymotion (44%) raises questions about the relevance of the initial moderation action on these services.

- Table showing complaints from service users

| *Table showing complaints from service* ***users*** | Internal appeals from users accessing the service from France | Results of internal appeals |
|---|---|---|
| *Bing* | Data not relevant | Data not relevant |
| *Dailymotion* | 6.46% (July-December) | 56% confirmation of initial decision (July-December) |
| *Google Search* | Data not provided | Data not provided |
| *LinkedIn* | 5 | 100% confirmation of initial decision |
| *Meta* (*Instagram* and *Facebook*) | 625 | 100% confirmation of initial decision |
| *Pinterest* | **Confidential** | **Confidential** |
| *Snapchat* | 65 | 100% confirmation of initial decision |
| *TikTok* | 34,404 (27,770 under the system dedicated to breaches of community rules and 6,634 under the "Report hateful content in France" system) | 40.8% of the initial decisions were overturned via the mechanism dedicated to breaches of community rules and 24.66% via the second mechanism |
| *Twitter* | 11,538 under the LCEN law | 600 were acted upon after appeals (action rate: 5.20%) |
| *Wikipedia* | *"Not applicable"* | *"Not applicable"* |
| *Yahoo portal* | No internal recourse | No internal feedback |
| *YouTube* | Data not provided | Data not provided |

## Working with experts in the fight against online hate: an asset for providers

Most platform providers have established, sometimes long ago, **working relationships with trusted flaggers**, both in France and the rest of the world. These relationships vary both in terms of the number and vitality of partnerships. Their effectiveness depends on the resources available to trusted flaggers and the investments they receive.

The DSA recognises **the key role played by these partnerships**, which contribute to a calmer digital environment. They will be strengthened and given a clear legal framework requiring platform providers to **prioritize** alerts from their partners. Providers and trusted flaggers will also be bound by **reciprocal transparency obligations,** which will be monitored by the national Digital Services Coordinators.

- Table showing collaboration with trusted flaggers in France

| Table showing the collaboration of services/operators with trusted reportersin France | Declared number of trusted reporters | Number of reports of hateful content declared under article 6-4 of the LCEN received in 2022 |
|---|---|---|
| *Bing* | None | Data not relevant |
| *Dailymotion* | *"It developed closer ties with several organizations that reported cases in 2022 to consider partnership arrangements. "* | 0 |
| *GoogleSearch* | "None. " | Data not relevant |
| *LinkedIn* | *"Does not currently have a formal programme for trusted reporters. "* | Data not relevant |
| *Meta (Instagram and Facebook)* | 17 (including four appointed) | 0 |
| *Pinterest* | **Confidential** | **confidential** |
| *Snapchat* | 4 | 0[15] |
| *TikTok* | 11 | 17 |
| *Twitter* | 5 | 242 (action rate: 66.12%) |
| *Wikipedia* | None | Data not relevant |
| *Yahoo Portal* | None | *"0 (N/A)"* |
| *YouTube* | 5 | Data not provided |

**Good cooperation between operators and the judicial and administrative authorities** is a guarantee of the effectiveness of the legal response to the dissemination of unlawful content.

The quality and maturity of the dialogue between providers and specialised investigative services is reflected in the **particularly high rate of** positive responses **(nearly 85%**) to requests for information from the French authorities for the purpose of identifying the potential author of hateful online content.

The DSA will further strengthen the teamwork's effectiveness: providers will be required to promptly respond to judicial or administrative authorities requesting them to take action against illegal content or to provide information about its author, and **to give specific reasons for any refusal to comply with the request**. Public annual reports will be written about these injunctions, the reasons they were issued and the providers' responses, contributing to the transparency of law enforcement actions against illegal online content.

[15] Modified on 27 july 2023 following a correction made by Snap to its declaration.

- Table showing requests and alerts issued by the French public authorities

| *Table showing requests and warnings issued by the* **French public** *authorities* | Number of requests and reports of hateful content within the meaning of article 6-4 of the LCEN received in 2022 | Rate of hateful content moderation action in 2022 | Number of suspected infractions forwarded to the competent public authorities in 2022 |
|---|---|---|---|
| *Bing* | 155 content removal requests | Data not provided | 0 |
| *Dailymotion* | 0 | Data not relevant | 0 |
| *Google* (*Search* and *YouTube*) | 6,017 requests for data received between January and 30 June | Information provided for 85% of requests | 0 |
| *LinkedIn* | 0 | Data not relevant | 0 |
| *Meta* (*Instagram* and *Facebook*) | 0 alerts but 25,451 requests for data (including 12,520 in the context of legal proceedings and 12,938 requests for emergency disclosure) | Data transmitted in around 85% of cases | Data not provided |
| *Pinterest* | **Confidential** | *"We carefully review each request and transmit the data for those that comply with the law and our policies. "* | **Confidential** |
| *Snapchat* | 0 | Data not relevant | Data not provided |
| *TikTok* | 11 | Data not provided | Data not provided |
| *Twitter* | 869 requests to remove content and 5,032 requests for information | 43.5% for deletion requests and 42.90% for information requests | 0 |
| *Wikipedia* | 0 | Data not relevant | Data not provided |
| *Yahoo portal* | 0 | Data not relevant | 0 |

## Methodology

The platforms covered by the law of 24 August 2021 posted an average of more than ten million single visitors per month in France over the last calendar year.

To prepare for this progress report, a questionnaire was sent on 28 April 2023 to the providers of the following thirteen services: *Google* (for *Google Search* and *YouTube*), *LinkedIn*, *Meta* (for *Facebook* and *Instagram*), *Microsoft* (for *Bing*), *Pinterest*, *Snap*, *TikTok*, *Twitter*, the *Wikimedia Foundation*, *Yahoo* (for *Yahoo Search*) and *Dailymotion*. With the exception of *Yahoo Search*, all these services are either established in France or are VLOPSEs as defined by the DSA.

In addition, between April and June 2023, Arcom observed the general terms and conditions and reporting mechanisms on each service to feed its analysis via the Google Chrome web browser for the services' website and via the iOS and Android operating systems for their application.

**For further reading:**

Publication Director:
Roch-Olivier Maistre