Arcom

Le régulateur de la communication audiovisuelle et numérique

Combating the dissemination of hate content online

Review of the resources deployed by online platforms in 2022 and future prospects

July 2023

Contents

Introduction

The development of information society services has profoundly transformed the way in which millions of users communicate and exchange information.

However, the increased use of these services, particularly platforms for sharing information between users (primarily social networks), has also given rise to new risks threatening the cohesion and democratic functioning of our society, such as the widespread dissemination of manifestly illegal content, the revelation of misinformation phenomena and, in some cases, disinformation, or the identification of harmful induced effects such as worsening public health problems, increasing conflict in online public debates, or the loss of confidence in our information spaces. This reality has led to expectations that digital service providers should become more responsible.

Given the urgency of the situation created by the distribution of illegal content online, some European countries – notably France, Germany and Austria – have anticipated this text by adopting the first binding national legislative frameworks ahead of time.

In France, the provisions of Article 42 of the Act of 24 August 2021 Consolidating Compliance with Principles of the French Republic, which introduced Article 6(4) into Act No. 2004-575 of 21 June 2004 on Confidence in the Digital Economy (LCEN), have stepped up the fight against hateful content by imposing procedural and resource obligations, both human and technological, on the main online platforms received in France. It entrusted the Audiovisual and Digital Communication Regulatory Authority (Arcom) with the task of supervising the implementation of these obligations. This system is largely based on certain provisions of the initial proposal for a Digital Services Act (DSA) presented by the European Commission in 2020[1]. These national provisions are due to expire on 31 December 2023[2].

In response to this expectation and drawing on the experience gained from national laws, particularly in France, Germany and Austria, the European Union (EU) legislator adopted the DSA on 19 October 2022[3] to establish harmonised rules for a secure, predictable and reliable online environment in which the fundamental rights of European citizens will be effectively protected.

[1] Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC.
[2] The Bill to Secure and Regulate the Digital Environment, tabled in the Senate on 10 May 2023, envisages extending this deadline to 17 February 2024.
[3] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services

To this end, the Regulation reaffirms and strengthens the system of limited liability for hosted content, but at the same time introduces a series of new obligations for all *"intermediary"*[4] service providers, as they are known, in terms of diligence, transparency, cooperation with public authorities, civil society and users, and moderation of illegal content.

From 17 February 2024, it will apply to all services concerned. From 25 August 2023, it will apply only to providers of very large online platforms and search engines (VLOPSEs)[5], some of which were designated as such by the European Commission on 25 April 2023[6].

At this stage of transition from a national to a European regulatory framework, Arcom has sought to exercise the prerogatives it derives from the Act of 24 August 2021 in order to anticipate the implementation of the DSA. As a result, it drew heavily on the latter when drafting its guidelines adopted in November 2022, and this report describes the implementation of these guidelines on the main platforms operating in France. The analyses carried out by Arcom in this review of the procedures and resources deployed by the platforms are based on observations and reports sent to it by the providers in response to an *ad hoc* questionnaire.

Arcom, which has been designated as the Digital Services Coordinator (DSC) under the DSA for France, alongside the French Data Protection Authority (CNIL) and the Directorate-General for Competition, Consumer Affairs and Prevention of Fraud (DGCCRF), which will be involved in implementing specific provisions of the DSA, in accordance with the Bill to Secure and Regulate the Digital Environment adopted on first reading by the Senate on 5 July 2023, intends to build on this initial experience to contribute to the implementation of enhanced regulation of online platforms, in particular social networks, and to the dialogue with European regulators on the implementation of the DSA, i.e., the European Commission, digital services coordinators and other competent authorities in the Member States.

[4] Intermediary services are those to which the DNA applies. Article 3 of the text defines them as information society services falling into one of these three categories:

i) **simple transport services,** where the activity consists of transferring information at the request of a third party or allowing third party access to the network (e.g., Internet service providers);
ii) **caching services,** which temporarily store information to facilitate subsequent transmission (e.g., content delivery networks or CDNs);
iii) **hosting services,** where the activity consists of storing data provided by a third party, including the specific category of online platform services (e.g., social networks).

[5] Online platform or search engine services with a monthly audience of more than 45 million active recipients in the EU, as designated by the European Commission.

[6] This list, which will be updated in the future, is available on the European Commission website: https://digital-strategy.ec.europa.eu/en/policies/dsa-vlops

I. Overview of Arcom's approach

A. Platforms concerned

The platforms covered by the Act of 24 August 2021 are those with an average of more than ten million unique visitors per month in France over the last calendar year. Additional obligations are imposed when the number of unique visitors exceeds fifteen million[7].

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*. All these services are VLOPSEs within the meaning of the DSA, with the exception of Dailymotion and *Yahoo Search*, which are not covered by this enhanced European regulatory regime.

The questionnaire referred particularly to the resource obligations set out in the DSA and invited providers to report problems they encountered in preparing to implement the Act and in applying French law.

B. Operator responses

Two providers, *Dailymotion* and *LinkedIn*, made a special effort to move quickly, submitting their contributions well in advance of the deadline set by Arcom. In addition, the former has opted to provide a detailed response, despite an audience, which – according to the data available to it – is below the threshold for triggering the obligations under national law.

Generally speaking, an examination of the platform reports again highlights four issues already encountered by Arcom[8].

- *Questions about the scope of the obligations*

Yahoo contested that it was subject to the provisions of Article 42 of the Act of 24 August 2021. According to its own data, *Yahoo Search*'s audience was below the threshold of 10 million unique monthly visitors on average in France in 2022. On the other hand, again according to the operator, the audience of the *Yahoo Portal* (an aggregate of *Yahoo Homepage*, *Yahoo News*, *Yahoo Style*, *Yahoo Sport* and *Yahoo Finance*) was above this threshold for the period and territory under consideration. The operator's contribution therefore concerns this second service. However, the operator considers that because of

[7] Thresholds defined by Decree No. 2022-32 of 14 January 2022 implementing Article 42 of Act No. 2021-1109 of 24 August 2021 Consolidating Compliance with the Principles of the French Republic. In accordance with article 3, *"only connections to a service, or to a separable part of a service, the main purpose of which is to classify, reference or share content posted online by third parties, shall be taken into account"*.
[8] In particular, as part of its annual reports on efforts to combat the manipulation of information on online platforms (available on the Arcom website).

its operating model, the *Yahoo Portal* is an editorial service and not an online platform. From this perspective, the aforementioned national law would not

apply. Since the declaration for this service is not useful, it has not been taken into account in this report.

- *Taking differences in platform models into account*

The *Wikimedia Foundation* considers the questionnaire to be ill-suited to the particularities of *Wikipedia's* collaborative and decentralised operating model.

Microsoft explains that *Bing* does not host user content or allow users to publish or share content on the service, and therefore believes that certain questions do not apply to it. However, the contribution sent by the operator provides interesting information, particularly in terms of collaboration with public authorities and their use of the reporting system.

- *Increased efforts by providers to improve transparency*

With the exception of *Pinterest*'s contribution, all the providers agreed to their responses being published in their entirety or almost in their entirety. However, this effort at transparency is sometimes to the detriment of the relevance of the information transmitted, particularly in the case of *Yahoo*. In contrast, *Dailymotion* and *Twitter* have made substantial contributions with a high level of transparency. In general, **a great deal of information was communicated publicly by the providers**, particularly on new subjects (e.g., the number, location, and working languages of their French-speaking moderators).

- *Difficulties encountered by providers in meeting deadlines*

The vast majority of the platforms questioned were unable to respond within a fortnight and asked for more time. In particular, contributions from *Google* and *TikTok* mentioned difficulties in gathering and communicating all the data requested in a timely manner, particularly figures – which may seem surprising given their status as VLOPSEs and their resources. Arcom believes that transparency is an essential element of the responsibility of online platforms in combating illegal content and behaviour and that, in this respect, the resources they are required to deploy to contribute to this effort must also relate to this transparency and their ability to provide this diligently.

Recommendation:

- ❖ strengthen appropriate resources, in particular the procedural, human and technological ones, to meet transparency obligations in a diligent manner.

C. Objectives and methodology of the additional observations carried out

In addition to the platforms' declarations, Arcom conducted observations on each service in support of its analysis, between April and June 2023, of the services' websites via the Google Chrome web browser and of their applications via the iOS and Android operating systems.

These comments related to the accessibility of the general terms and conditions[9] (GTC) of the services and the transparency of the moderation policies described therein, taking into account:

- accessibility of the GTC from any page of the service,
- a reference to the ban on posting hate content online,
- mention of the prohibition on abusing reporting systems,
- a description, in clear and precise terms, of the moderation mechanisms used in this area (procedures, means used, type of sanctions, mention of the means of redress, etc.),
- transparency regarding the procedures for suspending and terminating accounts that have repeatedly posted hate content online, and the level of clarity and precision of this information.

The reporting systems available to the public on the services were also examined, in particular:

- their accessibility (including by people not connected to an account on the service),
- ease of use,
- the accuracy and completeness of the headings of their reasons for considering hate content.

[9] As well as Community rules and transparency centres (or equivalents) for services.

II. Analysis of the resources used by online platforms to combat online hate

A. Transparency and clarity of the general terms and conditions (GTC) on the rules and conditions for applying moderation

On online platforms, users are required to comply with applicable national laws and any rules specific to the platform. If this is not the case, the platform may be required to moderate the offending content or account. In order to guarantee a *"secure, predictable and reliable"* online environment[10], French law requires providers to set out the rules and conditions of application of their terms and conditions of service in a clear and easily accessible manner.

Article 14 of the DSA (*"Terms and Conditions"*) requires all intermediary service providers to include in their terms and conditions *"information on any restrictions that they impose"* on recipients of the service, which must be set out in "*clear, plain, intelligible, user-friendly and unambiguous language*" and be publicly available "*in an easily accessible and machine-readable format"*. The clarity requirement is even more stringent when the service is primarily aimed at minors. In addition, VLOPSEs must in particular (i) publish their GTC *"in the official languages of all Member States in which they offer their services"* and (ii) provide recipients of their services with a *"summary of the terms and conditions"*.

Following the same logic, the French Act of 24 August 2021 stipulates that the *"general terms and conditions of use of the service"* must be easily accessible to the public and mention:

- a ban on posting illegal hate content online[11] ;
- the mechanisms for moderating such content, *"in clear and precise terms"*;
- measures taken in respect of users who have posted such content online and the domestic and legal remedies available to them;
- *"in clear and precise terms"*, the procedures for suspending or terminating the accounts of users who have repeatedly uploaded illegal content (if the platform implements such procedures).

Among the GTC of the services observed, only those of *Pinterest* explicitly mention French law, in a dedicated section at the bottom of the page which directs readers to the dedicated reporting form. National law is also mentioned in the *"Additional terms of use for Google Search*[12]*"*. Similarly, a redirect link to an article in the help centre dedicated to this legal text is provided at the foot of the *Instagram* GTC page. *Snapchat* mentions the prohibitions under French law on a specific page in its "*support*" area[13].

[10] Under the terms of Article 1 of the DSA.
[11] The scope of which is clearly defined by Article 6(I)(7) of the LCEN.
[12] NB: these are only valid for France.
[13] Sentence added in the revised version on 27 July 2023.

1. *The accessibility of GTC*

Three practices stand out in terms of the heading of the section allowing the user to access the service's GTC: the majority of the services observed (*Dailymotion*, *YouTube*, *LinkedIn*, *Facebook*, *Pinterest*, *Snapchat*, *Wikipedia* and *Twitter*) have opted for an explicit and logical reference to the terms "Terms and Conditions" or "*Terms of Use*"; *Google Search*, *Instagram* and *TikTok* use the more general term "*Conditions*", while *Bing* has chosen the less explicit term "*Legal*".

In light of observations made by Arcom, the degree of accessibility of the services' GTC appears satisfactory via a web browser on a computer: they are accessible in one *(Twitter*, *Instagram*, *Google Search*, *YouTube*, *Dailymotion*, *Wikipedia, TikTok* and *Bing*) or two clicks *(Facebook*, *LinkedIn* and *Pinterest*) from the service's home page. A maximum of two clicks is also generally required if the user is not logged in to an account on the service. However, the visibility of the section giving access to the service's GTC on this home page, as on the service's other pages, could be improved. This is particularly true for *YouTube*, whether the user is logged in to the service or not.

Access to the GTC is often slightly more complex on mobile via a web browser, with a maximum of four clicks required *(Facebook*, *Instagram* and *Twitter*).

The GTC are even less easy to access on the applications of certain services even though these applications are the primary means of accessing these services: Three clicks from the home page on the applications *Snapchat, LinkedIn* and *Facebook*, four on *Twitter, Bing* and *Pinterest*, and five on *Dailymotion*, *Instagram* and *TikTok*. What's more, the navigation path is often far from intuitive. Users will have to scroll down a long drop-down menu on *Snapchat*, *Facebook* and *TikTok*. On *Dailymotion*, you need to go to the "*Settings*" section. The same applies to *Twitter*, where you then have to click on "*Additional resources*". On *LinkedIn*, you need to go to the "Preferences" section and, on *Instagram*, to the "About" section. Finally, on *Pinterest*, you first need to click on the "*Saved*" section. To make it easier for users to understand the rules in force on these services, the headings could be more explicit and the path to accessing the GTC could be made more fluid.

Recommendation:

- ❖ make the path for users to access the general terms and conditions of service clearer, faster, and more fluid.

2. *The intelligibility of GTC*

Given the very nature of online hate content, offering intelligible GTC is both a challenge and a necessity.

Some providers suffer from the comparison in terms of the length of certain sentences *(Twitter* and *Snap*), lack of fluidity because of the French translation *(LinkedIn*), or the scarcity of examples provided *(Pinterest*).

In the vast majority of cases, the GTC are available in French, with the exception of those for *LinkedIn*, where certain sections are only available in English.

A good practice of some services is to provide summaries: on *LinkedIn*, the start of each section includes an insert and a light bulb icon, briefly and simply summarising the content of the section. On *Pinterest* and *TikTok*, at the end of each section there is a paragraph summarising the content. In the case of *TikTok*, however, the visibility of these paragraphs could be improved. These providers thus appear to have anticipated one of the obligations imposed by Article 14 DSA on very large online platforms, which requires VLOPSEs to provide recipients of their service with a *"summary of the terms and conditions"*.

For several departments, the layout of the GTC could be significantly improved to make them easier to understand. For example, the *Snapchat* page is particularly long and does not offer any means of navigation within it (cross-referencing between sections); the conditions applicable to users residing within the European Union appear after those applicable to the residents of the United States, although a geolocalised breakdown of the relevant conditions would prevent any confusion. For *Bing*, the page lists the GTC for all the many services operated by *Microsoft* and offers no cross-references. Finally, the layout of the GTC for the *Meta* services does not make them easy to understand (length, format not "justified", no cross-references to sections, etc.).

The GTC for *TikTok*, a service used extensively by minors, feature an uncluttered layout, relatively short syntax and non-technical vocabulary, which is in line with the need for greater intelligibility required by the DSA. Summaries of the content of each section in clear and easily understandable terms within the GTC of *TikTok* and *Pinterest* are also part of this logic.

Recommendation:

- ❖ improve legibility and navigation within and between the various pages devoted to general terms and conditions, Community rules, the help center (or equivalents), etc.

3. Information on restrictions

i) Mention of the ban on publishing hate content

There are two distinct practices adopted by the platforms:

- a clear and explicit mention of the ban on publishing hateful content in the GTC *(TikTok*, *YouTube*, *Dailymotion* and *Microsoft*);
- a concise and/or general mention (e.g., prohibition on publishing illegal content) in the GTC with reference to other pages (e.g., the service's community rules) where the prohibition on publishing hate content is specified (*Twitter*, *Facebook*, *Instagram*, *Snapchat*, *LinkedIn*, *Wikipedia*[14], *Pinterest* and *Google Search*).

[14] Readers are referred to the *"Universal Code of Conduct"*, available in English only.

With the exception of *Dailymotion*, *YouTube* and *Google's* "*Additional Terms of Use*", the providers'' GTC use a general statement (e.g., "*illegal content*") and/or restrictive (e.g., the term "*harassment*" is used to designate content of a hateful nature and does not explain the scope of the concept. It should be noted that the level of detail in the general terms and conditions of the service, in terms that are accessible to all, determines whether the rules in force are properly understood by those to whom they are addressed.

Recommendation:

- clearly specify, within the general terms and conditions of the service, what content and behaviour are prohibited by national law and the operator's rules, and in particular the prohibition of incitement to hatred and online harassment.

ii) Description of hate content moderation policies and mechanisms

The GTC of the services observed do not mention any moderation policies specific to hate content, but deal in general with the moderation of harmful and/or illegal content. There are a number of different practices in this area.

The GTC of *Pinterest*, *LinkedIn* and *Bing* provide minimal explanations, a few lines outlining their approach in clear terms (e.g., they list certain content that may be subject to a moderation decision and mention certain sanctions) and invite the reader to look up related pages (e.g., help centre or community rules) to find out more about the subject.

While they may be more detailed, those of *Twitter* could be clarified and illustrated. For example, less easily understood reasons for moderation is worded as follows: *"Our provision of services to you is no longer commercially viable"*[15].

The GTC of *Snap*, *Meta*, *TikTok* and *Google* set out more fully and precisely the procedure that may lead the operator to suspend or delete content or an account. These pages give a more or less brief description of the systems, measures and procedures in force on the platform, sometimes with references to appendices setting out the platform's policies.

In the case of *Snap*, the description of the moderation methods used, in particular to detect and examine content, remains evasive and would benefit from being more detailed. Conversely, the GTG of the *Meta*, *Google* and *TikTok* services explicitly mention the use of automated tools for moderation purposes. *TikTok* explains that its approach is based on a combination of automated tools, human moderation, and user reports.

Because of the collaborative and decentralised approach of its moderation policy, comparing *Wikipedia* with other services is hardly relevant. However, the description of this policy in the platform's general terms and conditions would benefit from greater accuracy and transparency.

[15] "4. Use of the Services", section "Termination of these Terms and Conditions". Source: https://twitter.com/fr/tos

Dailymotion is demonstrating a significant degree of transparency. An appendix to its GTC entitled "*Policy on Prohibited Content*" sets out (i) the various categories of prohibited content (e.g., "*D. Hate Content*"), (ii) the detection and reporting of such content (automatic detection devices, dedicated reporting tool, reporting by email or post), and (iii) the consequences of non-compliance with the policy on prohibited content. This last part first presents the moderation actions applicable to prohibited content (affecting its availability, visibility, accessibility, or monetisation) before explaining the process for appealing a moderation decision. The high level of transparency and didactic effort seems likely to ensure that users understand the rules in force.

On most of the services observed, the possibility of accessing an internal system for handling complaints about a moderation decision is mentioned in the GTC *(Twitter*, *TikTok*, *Facebook*, *Instagram*, etc.), *Google Search*, *YouTube*, *Dailymotion*, *Wikimedia Foundation*), in community rules *(LinkedIn*) in the "*assistance*" area (*Snapchat)* [16]. According to these GTC, complaints may relate to account moderation decisions (*Twitter* and *Google Search*) or content moderation decisions *(Wikipedia*, *Dailymotion* and *Instagram*), or both *(TikTok*, *YouTube*, *Facebook* and *LinkedIn* via its "Professional Community Policies").

Among these services, and with the exception of *Wikipedia* and *Dailymotion*, the GTC reserve this right of complaint for "positive" moderation decisions (e.g., suspension or deletion of an account or content), excluding decisions not to act on a report.

However, two platforms are notable exceptions:

- *Pinterest*: the possibility of appealing a moderation decision is not clearly mentioned in the GTC. You need to contact the help centre to find information on this subject.
- *Bing*: no mention of the option to appeal against a moderation decision is made in the GTC.

This practice does not seem to comply with the provisions of Article 14 DSA, which stipulates that the GTC must provide information on *"the rules of procedure of their internal complaint handling system"*.

In addition, the dedicated sections of the GTC of *Google Search*, *YouTube*, *Facebook*, *Instagram*, *Twitter* and *LinkedIn* (via their "*Professional Community Policies*") refer the reader to a help page on how to obtain redress. The same applies to *Wikipedia*, but the page in question is available only in English. Finally, in *TikTok's* GTC, the section on "TikTok Rights", which addresses this subject, includes a link to a contact form enabling anyone to appeal against a moderation decision.

In general, the internal systems for handling complaints about a moderation decision are accessible (i) after complex navigation within the GTC and/or help pages of the service, (ii) via a dedicated Internet request (e.g., "*call for suspension/deletion of account/content [name of service]*") or (iii) by sending a message to the user who has been the subject of the moderation decision[17].

[16] Revised on 27 July 2023: a previous version of the report did not mention that Snapchat provided information about the possibility of appealing a moderation decision in its "support" area.
[17] For an analysis of the remedies available to a recipient of the service, please see II. D).

Finally, the GTC of *Dailymotion* and *YouTube* and the "Additional Terms of Use of *Google Search*" explicitly mention legal remedies, which seems likely to ensure that users are fully informed.

In addition, the GTC of certain services explicitly mention the right of their providers to notify illegal content to a third party, including the police (*Snapchat*, *Facebook* and *Instagram*) and the competent legal authorities (*Dailymotion* and *Wikipedia*).

Recommendation:

- be as clear and explicit as possible, within the general terms and conditions, about the existence and operation of the internal mechanism for appealing moderation decisions, both for people who report content and for those whose content is moderated.

B. Hate content reporting systems

Online platform providers are not subject to a general obligation to monitor the content published on their service, but they are required to remove any content that is clearly unlawful once it has been reported to them. To ensure effective moderation, the law requires providers to provide a reporting tool designed to facilitate the act of reporting to the greatest possible extent.

Article 16 of the DSA provides that *"providers of hosting services shall put mechanisms in place to allow any individual or entity to notify them of the presence on their service of specific items of information that the individual or entity considers to be illegal content"*. These mechanisms must be *"easy to access and user-friendly"* and help to facilitate the submission of *"sufficiently precise and adequately substantiated"* notices.

With the same objective in mind, the French Act of 24 August 2021 requires platforms to put in place " *an easily accessible and user-friendly system enabling any person*" to bring to their attention, *"by electronic means"*, content considered to be hateful in nature.

Two services stand out in this respect.

Because of the specific features of its model, *Wikipedia* does not have a "traditional" reporting system. When confronted with problematic content, users can (i) correct it themselves by modifying the page, (ii) explain the problem on the discussion page for the article concerned, or (iii) ask for help on the "*New Users Forum*"[18]. The platform's moderation architecture is remarkable in that it relies primarily on users, while ensuring stable expertise in moderating illegal content thanks to *Wikipedia*'s "*administrator*" users[19].

On *Bing*, a user wishing to report illegal content is redirected to the form offered by the association Point de Contact on its own website[20]. The disadvantage of this system is that the platform operator's moderation teams cannot be contacted directly. While a

[18] Source: https://fr.wikipedia.org/wiki/Aide:Accueil/Signaler
[19] Source: https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Administrateur
[20] Source : https://www.pointdecontact.net/cliquez-signalez/

reporting form is available from *Microsoft* it is not easily accessible[21] and, above all, does not include any grounds for reporting hate content.

Furthermore, among the services observed, only the GTC of *Wikipedia*, *Dailymotion*, *YouTube*, *Facebook*, *Instagram* and the "*Additional Terms of Use of Google Search*" explicitly state that abuse of this reporting tool is prohibited.

1. *Are the reporting forms easy for users to access?*

i) For a user not logged in to an account on the service

French law stipulates that the service's reporting system must be accessible *"to anyone"*. This includes users who are not logged in to an account, when the service can be accessed without registering and logging in. It should be noted that this obligation raises a specific technical issue for providers: as the sender of the report is not connected to the platform, the operator does not know the sender's contact details and must therefore deploy a specific system to receive these reports under the conditions required by law (i.e., by acknowledging receipt to the user and informing him/her of the action taken on the report).

Google Search, *YouTube* and *TikTok* offer an *ad hoc* form allowing anyone to report hateful content under Article 6(4) LCEN[22]. Arcom emphasises the value of this practice but considers that the accessibility of these forms could be improved insofar as (i) many clicks are required to access the *Google Search* form, (ii) the *YouTube* form can only be accessed after navigating through the help pages and (iii) the *TikTok* site has loading problems that prevent it from being displayed.

In this respect, Arcom considers that the existence of forms specifically dedicated to Article 6(4) LCEN should not lead the platform to consider, in its transparency reports, that only content reported via these forms should be counted as hate content reported to the service in France. Not everyone is aware of the existence of these forms, and their title may discourage users. Therefore, those reported under the rules in force on the platform via the content reporting system should also be taken into account.

The services operated by *Meta* also have a form dedicated to reports under the LCEN. However, (i) it can only be accessed after complex and not very explicit navigation through the integrated reporting system followed by the service's help pages or a very specific Internet query (e.g., "*formulaire signalement respect république Facebook*" ("*Facebook Republic Compliance Report Form*")) and (ii) it is partially drafted in English. It could be improved to make it more accessible to the public.

Modified during 2022, the reporting system on *Dailymotion* now allows users who are not logged in to report content. *Snapchat*, for its part, seems to have an *ad hoc* form allowing anyone to report "*abuse*". However, during an observation carried out in June 2023, the page was inaccessible.

[21] Only accessible via a dedicated internet request ("Bing report") and several clicks. On the other hand, the query "Bing reporting form" does not, counter-intuitively, give access to it.
[22] Resulting from the Act of 24 August 2021 – please see note 5.

Finally, *Twitter*, *Pinterest* (except via an *ad hoc* form) and *LinkedIn* do not offer a facility for offline users to report illegal content.

Recommendation:

- ❖ in transparency reports, take into account:
  - hate content reported via *ad hoc* forms and via the content reporting system;
  - those removed (i) based on the general terms and conditions and (ii) under national law.

ii) For a user connected to an account on the service

All the services observed have reporting systems available to users with an account. These tools can be accessed in less than two clicks from any content on the service[23]. They generally allow hate content to be reported in less than five clicks, with the exception of *Twitter* and *Google Search* which require more clicks.

However, on many services *(YouTube*, *LinkedIn*, *Facebook*, *Instagram*, *Pinterest*, *Twitter*, *TikTok*), access to the reporting system is conditional on clicking on a button with an obscure title (e.g., "...", please see screen shot below).

Screen shot of the Twitter application, 18 July 2023 (source: Arcom)

The accessibility of these systems should therefore be improved. *Snapchat* also uses headings that are sometimes not very explicit: for example, a user wishing to report a 'friend' profile because of illicit behaviour will first have to click on the "*Manage friendship*" heading. On the other hand, the button for reporting content is clearly visible and explicit on *Dailymotion*.

[23] There are a few exceptions; for example, it is not possible to report an 'Event' on Facebook.

Pinterest has a dedicated reporting form that allows any connected user to report content under the LCEN. However, the title of the reason for accessing it, "*the Act Consolidating Compliance with the Principles of the French Republic*", says very little to the public. In the interests of clarity, this title should be made more explicit.

iii) Special case of reporting an account on video-sharing platforms

The accessibility of the system for reporting accounts on *YouTube* could be made much easier. Indeed, on an account page, the idea of going to the last section, entitled "About", does not seem intuitive. What's more, access to the reporting system is only available via a small black and white flag icon.

On *Dailymotion*, it is not possible to report an account.

Recommendation:

- improve the accessibility of reporting systems, in particular by using more explicit symbols and headings.

2. Is the heading related to reasons for reporting clear?

The degree of clarity of the reasons for reporting hate content varies from one service to another.

It is high when the service has a reporting form and/or a reporting category dedicated to hate content under the LCEN. This is the case for *Google Search, YouTube*, *Pinterest*, *Facebook*, *Instagram* and *TikTok*. In addition, *Pinterest*, *Google Search* and *YouTube* are distinguished by their concise headings, whereas the longer headings on the dedicated form accessible via the *Meta* services indicate the legal provision to which they correspond. In addition, two reasons for *Google Search* and *YouTube* are illustrated with examples. The first sends the user back to

to the text of the Act and the second, when indicating the reason for the report, to a page dedicated to reporting hate content under the LCEN.

As for the mechanisms for reporting hate content integrated into the *Dailymotion*, *Meta*, *YouTube* and *LinkedIn* services, they are clearly designed. The proposed reasons are also illustrated with examples. *Snapchat's* are easy to understand: the list of reasons for reporting is long, but the reasons are classified into subsets to make them easier to understand.

As the scope of hate content under French law encompasses different types of offence, platforms have opted to cover some of these offences under different headings, for the sake of clarity for users.

When it comes to *Twitter*, on the other hand, the distinction between headings for the different reasons for reporting content is not always intuitive. By way of example, hate content seems to correspond to the following two sub-reasons at the same time: "*The user*

threatens to use violence or hurt someone" and "*The comments incite hatred towards a protected category"*. This ambiguity may discourage users from reporting.

As for *Google Search*, it should first be noted that the presentation of the step following selection of the Google product to which the report relates is confusing: the user is asked to tick "*Yes*" or "*No*" without being asked any questions. However, the answer will have an impact on the rest of the process by allowing access, or not, to the form devoted to "*non-legal reasons linked to the regulation for reporting content*". In addition, in the latter form, the reason "*Doxxing: report content featuring your contact details and containing explicit or implicit threats, or explicit or implicit incitements to action intended to harm or harass*" could be made more explicit.

Finally, the majority of providers *(Dailymotion*, *LinkedIn*, *Meta*, *Microsoft*, *Pinterest*, *Snap*, *TikTok* and *Twitter*) did not report any particular difficulties in including online hate-related reasons in their reporting systems. On the other hand, *Google* has reiterated the need to strike a balance between simplicity of use of its reporting system and the implementation of legally compliant solutions.

However, it would appear that the wording of the reasons for reporting could sometimes be improved to cover more explicitly the scope of national law. For example, while all the services include a reason covering bullying in general, none explicitly specifies that this may include bullying at school[24]. However, it should be noted that *Dailymotion* specifies that *"child abuse"* includes online harassment, and that *Facebook* is unique in that it sends a message that appears when a report of harassment is made, which specifically mentions the enhanced protection offered by its rules with regard to minors. Twitter, on the other hand, has a reporting system where users have to click on the first heading (*"The comments made are inappropriate or dangerous"*) to access the harassment section.

3. Some good practices

i) Across all the services observed

A frequently observed good practice is to spell out and illustrate the headings of the reporting system (e.g., *Dailymotion*, *Google Search*, *LinkedIn*, *Facebook*, *Instagram* and *Pinterest*) and a brief reminder of the rules in force regarding content (e.g., *Google Search*, *Facebook*, *Instagram*, *Pinterest*, *Snapchat* and *TikTok*) before the user completes their report.

It may be considered good practice for users to be able to add a description of any element they consider useful when analysing their posting *(Google*, *YouTube*, *Pinterest, Snapchat*[25] and *Twitter*). On the other hand, the potential deterrent effect of the practice observed on *Dailymotion* of making such a description compulsory, even if only briefly.

[24] School bullying has been an offence under the Criminal Code (Article 222(33)(2)(3)) since Act No. 2022-299 of 2 March 2022 to Combat Bullying at School, which also included it among the reasons for online hatred within the meaning of the LCEN.
[25] Revised on 27 July 2023: a previous version of the report stated that the description was mandatory on Snapchat, not optional.

The reporting forms dedicated to Article 6(4) LCEN (*Google Search*, *YouTube*, *Pinterest*, *Facebook*, *Instagram* and *TikTok*) also offer this feature, on an optional basis with the exception of *YouTube*, thus anticipating one of the provisions of Article 16 DSA, which stipulates that notifications issued via the notification system must contain *"a sufficiently substantiated explanation of the reasons why the individual or entity alleges the information in question to be illegal content"*.

The possibility offered by *Pinterest* for users who report content using the form dedicated to Article 6(4) LCEN to indicate which part(s) of the content their report concerns (e.g., the use of the *"Profile image"*, *"Profile name"*, *"Profile description"* or *"Other..."* option will improve the accuracy of the information transmitted.

The feature offered by *LinkedIn*, allowing users to indicate whether they would like to receive updates on the status of their report, is also good practice, as recommended by Arcom in its guidelines.

Finally, by allowing a user to report several tweets from the same account via a single report, *Twitter* stands out with a practice that seems likely to streamline the reporting process.

ii) Special case of video content sharing platforms

Dailymotion allows any user making a report, whether or not they are connected to the service, to indicate the time stamp from which, in their opinion, it would be appropriate to concentrate the analysis of the video content. *YouTube* offers the same feature. However, on this platform, reports can only be made by a user logged on to the service. This feature would be even more relevant if it allowed the user issuing the report to indicate a complete time interval.

Finally, another good practice is the option offered by *YouTube* to indicate whether their report also applies to links included in the description of the video reported.

Recommendations:

- illustrate the headings of the reasons for reporting with specific examples;
- provide a brief reminder of the rules in force on the service with regard to content before users finalise their reports;
- offer users the option to accompany their report a description of any element they consider useful for analysing it;
- offer the user the option to indicate the part(s) of the content that their report concerns;
- allow users to report several items of content from the same account in a single report;
- allow the user reporting video content to indicate a complete time interval and not just a start time stamp;
- offer users the option to indicate whether their report also applies to links included in the description of the content reported.

Furthermore, in keeping with what it already indicated in its guidelines, Arcom reiterates the following recommendation:

- allow users to indicate whether they wish to be kept informed of the progress of their report.

4. Combating abusive reporting

Article 23 DSA will require hosting service providers to take measures to combat the misuse of illegal content notification systems. Article 6(4) LCEN provides that platforms may temporarily suspend or, in the most serious cases, permanently suspend users who report abusive content.

These provisions are designed to protect users from fraudulent practices, taking the form of massive *"raids"* of reports with the sole aim of deceiving providers' moderation by unjustifiably suspending the targeted account.
However, the providers interviewed reported having taken few preventive or coercive measures to discourage misuse of their reporting tools. *LinkedIn* or *Snapchat*, for example, presume that any content reported was reported in good faith.

Twitter and *Meta* state that they are able to suspend users who misuse the reporting tools, but have decided not to in 2022.

Recommendation:

- ❖ be vigilant and, where necessary, maintain the capacity to react to misuse of reporting tools.

C. Resources used by online platforms to moderate hate content

Providers must treat the reports they receive in a diligent manner, implementing the appropriate resources, particularly in terms of moderators.
They are also required to pay particular attention to notifications from *"trusted flaggers"*, entities recognised for their expertise in combating illegal content online.

Article 16 DSA will require hosting service providers, in particular those offering online platform services, to put in place notification and action mechanisms enabling any individual or entity to report illegal content within their service. These providers will have to ensure that these notifications and the illegal content they relate to are dealt with "*in a timely, diligent, non-arbitrary and objective manner"*.

Article 22 of the Regulation reinforces this diligence requirement with regard to notifications of illegal content sent by entities designated as "*trusted flaggers*", which must be given priority treatment. Hosting service providers are required to implement appropriate *"technical and organisational measures"* to respond *"without undue delay"* to notifications sent by trusted flaggers, who therefore have a special role to play in analysing and combating the public dissemination of illegal content online.

It also strengthens the procedural obligations aimed at guaranteeing the diligent processing of notifications sent by *"trusted third parties"*, recognised as such based on their *"particular expertise and competence detecting, identifying and reporting illegal content"*, provided that they represent *"collective interests"* and offer *"guarantees of diligence and objectivity"*.

1. Human resources and procedures implemented to deal with user reports

Arcom asked providers about the human resources deployed to process reports.

The Authority notes that a significant proportion of providers[26] still refuse to publicly disclose the number of people dedicated to this task.

However, *Twitter* claims to have 149 moderators *"including French speakers"*[27], *Dailymotion* estimates that it will have "*around thirty moderators*" by 2022, while *LinkedIn* states that its internal team is made up of "*around a hundred reviewers*", including "*around 29 French speakers*". The *Wikimedia Foundation* estimates that *"around fifty"* volunteer contributors make a particular contribution to combating illegal content on their platform, while pointing out that the specific principles of collaborative moderation at the online encyclopaedia make this figure of little relevance.

Finally, the Authority notes that providers who are transparent about their moderation mechanisms specify that they subject all user reports to a twofold check, firstly, with regard to their general terms and conditions or standards of use (*Meta, Twitter, TikTok*) and, secondly, with regard to the legal provisions applicable in France.

In the event that the content reported does not contravene community standards or the operator's internal policies, but is nonetheless illegal under French law, the content in question will be removed for France only.

Meta reports that of the 4,807 items of illegal content that it claims to have blocked in France in 2022 on its two platforms, only 19 were blocked solely because of a breach of the provisions of Article 6(4) LCEN. This operator gives clear priority to removals for breaches of community standards. While this method of examination is not questionable, the Authority notes that it has the effect of minimising, in the figures made public, the number of illegal hate contents that are actually moderated.

Recommendations:

- ❖ increase the transparency of moderation policies by making public the number, working language and location of moderators employed by the operator;
- ❖ ensure that the human resources dedicated to moderating illegal content are adequately sized.

2. Recognition of trusted third parties in France

In the light of their responses, providers may be classified into two distinct groups: those who have established more or less close links with a set of trusted third parties *(Twitter, Meta, Google, TikTok, Snapchat*) and those who say they do not work with trusted third

[26] *Google*, *Meta*, *Snapchat*, *Pinterest*, *Yahoo* and *TikTok*

[27] Interviewed by Arcom, *Twitter* clarified that this was the number of moderators for Europe, responsible for dealing with breaches of the platform's rules or local laws. This figure may seem low given the size of the platform and the liveliness of the exchanges that take place on it, but it is difficult to assess without being able to compare it with the resources deployed by the other major online platforms, which have not had the transparency of *Twitter* in their response to Arcom. The application of the DSA, which requires completeness and transparency on the part of the platforms, will make it possible to carry out a solid comparative exercise.

parties because of the nature of their platform *(Microsoft* for *Bing*, *Yahoo*, *Wikimedia Foundation*) or for internal policy reasons (*Pinterest*, *LinkedIn*).

Dailymotion states that it has had *"long-standing, albeit now severed, links"* with unspecified trusted flaggers, but that it is willing to restore these links. The operator states that it will set up a reporting channel dedicated to trusted third parties during 2022.

3. Working with trusted third parties

Among the providers who say they work with trusted flaggers based in France, the number of partners varies widely. Furthermore, there is no correlation between the number of trusted reporting partners declared and the rate of content reporting. In addition, some providers were unable to quantify the number of reports transmitted by their trusted partners.

Meta, which declares 17 *"trusted partners"* in France, does not specify the number of reports of illicit content from them, while *TikTok* (11 recognised *"trusted flaggers"*) claims to have received 17 reports of hate content within the meaning of the LCEN for the year 2022 from its partners. Conversely, *Snapchat,* which says it work with 4 *"trusted third parties"*, reports that it has not received any reports of hate content from them[28].

Similarly, *Twitter*, which acknowledges five trusted flaggers in France, reports having received 242 reports of hate content in 2022, and having removed the content reported in 66% of cases. This figure is high compared to the 24% and 43% of removal decisions taken by the operator following a user report and a law enforcement report respectively.

Twitter states that it has a reporting channel in the form of a portal dedicated exclusively to trusted flaggers, which appears to be good practice in terms of guaranteeing the traceability of reports and compliance with the requirements of the DSA regarding the prioritisation of reports sent by trusted flaggers.

In contrast, some providers do not mention the existence of a specific reporting system to ensure that these particular notifications are handled appropriately.

This is the case with *Meta* and *Google*, the latter specifying that it is not in a position to provide figures in relation to the number of reports sent by them.

Finally, Arcom notes that the providers have not provided any data enabling it to objectively assess their speed in processing notifications from trusted flaggers. It highlights the fact that this is a substantial obligation of the DSA and a guarantee of the effectiveness of the system for combating the dissemination of illegal content.

Recommendations:

- ❖ systematise, in transparency reports, the distinction between the origin of the report between users, trusted flaggers, and public authorities;
- ❖ collect and make public the data needed to objectively assess the speed with which notifications sent by trusted flaggers are processed.

[28] Revised on 27 July 2023 following a correction made by Snapchat to its declaration.

D. Reviews procedures

Providers of online platforms are required to put in place a system enabling anyone to challenge a moderation decision taken by the operator. Such a system must be easy to access and use by all users.

The DSA requires all providers of hosting services, particularly online platforms, to provide clear and precise reasons for any restriction on use against a recipient of the service. These restrictions, listed in Article 17 of the Regulation, may in particular consist of the removal of unlawful content or the suspension or deletion of an account.

However, these decisions must be open to challenge by the recipients of the service. Article 20 of the Regulation stipulates that hosting service providers must set up an internal complaint handling system allowing appeals against decisions within a reasonable timeframe and guaranteeing non-discriminatory and non-arbitrary treatment.

Article 6(4) LCEN consolidates the obligations on major online platform providers to provide reasons for decisions to remove content or suspend an account. The existence of redress procedures must be clearly set out in the general terms and conditions of use, and reasons must be provided for decisions. Penalties such as suspension or deletion of a user account must be proportionate to the offence and subject to appeal.

The rate of remedies sought against initial decisions varies from one platform to another, but is low overall. Providers rarely distinguish between remedies against content decisions and remedies against account suspensions or deletions, even though the infringement of users' freedom of expression is more serious in the latter case.

In addition, the rate of reversal of this decision is generally quite low. *LinkedIn* claims that of the 3,064 pieces of hate content removed in 2022, only five were appealed, and none of the decisions appealed against were reversed

Twitter reports that of the 11,538 appeals received under the LCEN in 2022, 600 were the subject of action following a review, i.e., 5.2%. The nature of these actions (restoration of the content or account, or conversely, deletion of content previously considered lawful) is not detailed.

TikTok stands out for its high number of appeals (27,770 applications) and a particularly high rate of reversal of decisions (40%). The figure is similar for *Dailymotion* (44%), although the number of initial decisions is much lower. These very high rates of review of decisions demonstrate, if any proof were needed, the usefulness of redress procedures, but above all they raise questions about the relevance of the initial moderation.

Some platforms consider that the remedies sought by users have little added value; *Meta*, which noted 625 requests for review in 2022, explains that no decision has been reversed by its services because of the *"poor quality and lack of legal grounds"* of the requests for review, while deploring the presence of a certain amount of spam. However, the Authority notes that it is the responsibility of service providers to ensure that mechanisms for challenging moderation decisions are *"easily accessible and user-friendly"*.

Finally, we note that none of the providers has reported any legal action taken against its decisions.

Recommendations:

- allow any user to contest a moderation decision and ensure that these appeals are treated equally, without any particular legal argument being required;
- when the rate of reversals, following an appeal, of moderation decisions or actions is high (as is the case for *TikTok* and *Dailymotion*), take appropriate measures to assess the relevance of the initial moderation and remedy it where necessary.

E. Duty to cooperate with national judicial and administrative authorities

Online platform providers are required to put in place procedures and resources, both human and technological, to ensure a rapid response to requests from law enforcement agencies and the judicial authorities.

These requests may relate to illegal content within the meaning of French law or to data enabling a user of the service suspected of having distributed illegal content to be identified.

The DSA requires all intermediary service providers to respond to orders issued by the competent national judicial and administrative authorities based on Union or national law *"without undue delay"* and to inform them of the action taken in response to these orders.

The judicial or administrative authorities may order the service provider to take action against illegal content or to transmit information about a recipient of the service, under Articles 9 and 10 of the Regulation.

In addition, intermediary service providers, while not subject to a general obligation to monitor or actively seek out illicit acts or activities on their services, are under an obligation to report suspected criminal offences to judicial or law enforcement authorities where there is reason to fear a threat to the life or safety of third parties.

Article 6(4) LCEN imposes such a diligence obligation by requiring providers of online platforms to implement proportionate procedures and human and technological resources to enable:

i) the judicial or administrative authorities to be informed without undue delay of the action taken following receipt of an order concerning hate content on their service;
ii) receipt to be acknowledged without delay of requests from the judicial or administrative authorities for any data in their possession that may enable them to identify users who have posted illegal content online.

1. Procedures and human and technical resources to enable requests from public authorities to be processed diligently

The ability of online platform providers to receive and respond promptly to requests from the public authorities plays a major role in the effectiveness of combating the dissemination of hate content online.

The majority of providers, in this case *Google*, *Meta*, *Snapchat*, *Yahoo*, *Microsoft (Bing)*, *LinkedIn* and *Twitter* say they have set up a communication channel dedicated exclusively to receiving requests from law enforcement and judicial authorities. This practice seems particularly useful for guaranteeing the traceability of requests and the authentication of applicants.

Google specifies that investigating departments using the dedicated interface systematically receive an acknowledgement of receipt as soon as the request is sent, and can track the progress of requests by logging on to their digital file.

Providers who have set up a specific contact protocol for law enforcement agencies state that the requests they receive through this channel are analysed by teams of moderators, generally made up of lawyers, to examine the legality of requests prior to any response.

While the majority of providers claim to deal diligently with requests sent to them via the communication interfaces reserved for law enforcement agencies, *Snapchat* is unique in claiming to deal with requests relating to the most serious offences, such as imminent threat to life or limb, within 30 minutes.

Finally, the *Wikimedia Foundation, Dailymotion* and *TikTok* state that they have set up a contact address enabling the investigating authorities to forward their requests.

Recommendation:

- ❖ allocate a number of analysts commensurate with the need to process requests from public authorities promptly.

2. Reporting suspected criminal offences to the law enforcement authorities

No operator has reported a suspected criminal offence to the French authorities during 2022.

However, *Snapchat* states that it is proactively engaged with the security forces of the countries in which it operates and claims to promptly report any situation that appears to present a particularly significant risk to third parties (such as a bomb threat or a threat of attack) to federal security agencies, when the estimated danger is located

within the United States, or to Interpol, when the threat concerns another jurisdiction.

3. *Receiving and processing orders from the French authorities*

The requests that the French authorities have been able to transmit to providers under article 6(4) LCEN are, on the one hand, orders relating to hate content and, on the other, requests for information aimed at identifying the authors of hate content.

The responses from providers are particularly varied in this respect, some being able to identify precisely the number and nature of requests received concerning illegal content in general, or even hate content more specifically, as well as the actions taken following these referrals, while others do not distinguish between reports of illegal content sent by users and requests for information sent by the French investigating authorities in 2022.

i) Reporting of illegal content by the French authorities

Few providers claim to have received reports from the French authorities, and when they do, the data varies greatly from one platform to another.

Twitter reports that it has received a large number of requests from the French authorities to remove content (869). As mentioned above, the operator reports a removal decision rate of 43% following these requests. This relatively low level is surprising, particularly given the rate of action taken in response to alerts from trusted flaggers (66%). Arcom invites *Twitter* to examine the reasons for this.

TikTok only recorded 11 reports for 2022. Whereas *Meta* notes that no content has been reported to it by law enforcement agencies, *Bing* (which received 155 requests to remove content) reports that the 155 reports of hateful content it received came solely from French *"government authorities"*, not from users.

Among the providers who received a large number of reports of hate content, *Google* (which received 805 content reports through *Google Search* and 2,355 reports through *YouTube* solely via its forms dedicated to article 6(4) LCEN) claims that it is unable to distinguish between content reported by users and content reported by law enforcement agencies. The Authority notes that the transparency obligations set out in Article 15 DSA will require providers to make public in an annual report the number of orders received from the authorities of the Member States, broken down by type of illegal content concerned, the Member State which issued the order, and the average time taken to inform the issuing authority of its receipt and act on the order.

Finally, several providers claim not to have received any requests from the public authorities to remove hateful content, including *Meta*, *LinkedIn*, *Snapchat*, *Yahoo* and the *Wikimedia Foundation*.

ii) Information requests

The requests for information issued by French judicial or administrative authorities to providers mainly concern the platforms most targeted by requests to remove content.

The providers contacted report no particular difficulties in responding to requests from the judicial and administrative authorities.

The platforms most targeted by these requests, *Meta* and *Google*[29] in particular, claim to respond favourably in the majority of cases, with the rate of positive responses exceeding 80%, while the providers of the *Yahoo*, *Bing* and the *Wikimedia Foundation* claim not to have received this type of request.

Conversely, we note that providers such as *Snapchat* or *TikTok* say that they are unable to quantify the number of requests for information received.

III. Perspectives

A. Providers gradually taking on board their social responsibility

The major online platforms, in particular the largest social networks and search engines, play an essential role in access to information, participation in public debate and its preservation; as new *fora*, they contribute fully to the vitality of a democratic society.

However, their business model is largely based on advertising, and therefore on the unlimited exploitation of users' attention; their operating architecture is based on the individualisation of the content presented (including advertising) using massive algorithmic processing of the individual data collected. These two characteristics of the very large online platforms and search engines entail systemic risks for our societies.

There is therefore a risk that these platforms or search engines will artificially increase the visibility of the most controversial comments, facilitate the dissemination of manifestly illegal content, or allow the viral propagation of harmful content, all of which are likely to increase the conflict in the public arena and undermine social cohesion.

Moreover, well-known public health issues (e.g., risk of obesity, anorexia, behavioural disorders, addiction to digital services) may be magnified by the new information dynamics resulting from the use of the largest platforms or search engines.

We need to identify upstream, assess downstream, mitigate and, more generally, combat these systemic risks to our societies at the very design stage of digital services in order to preserve these new areas for the exercise of freedom of expression as a common good.

These phenomena have now been clearly identified.

This report illustrates the fact that a growing number of providers, aware of the impact of their services on the functioning of our democratic societies, have devised, often with some success, solutions to mitigate the most obviously damaging misuses of their services (dissemination of child pornography or terrorist content, sales of illegal goods and products, incitement or provocation to damage property) and to anticipate and prevent the the most damaging effects on our social dynamics, unintended yet real, and which constitute a challenge to our fundamental rights.

[29] Google states that it received 6,017 requests for data from the French authorities between January and June 2022, and that it responded favourably in 85% of cases.

These measures have, however, developed under a system of self-regulation, with indirect and informal pressure from civil society players or governments, and are now proving insufficient or imperfect to meet the social and democratic challenges.

Implementing a voluntary approach to increased accountability, designing services that are intrinsically safer for users and our societies, anticipating harmful side-effects and developing mitigation mechanisms all come at a cost that can be significant and run counter to the economic rationality of platforms whose model generally depends on maximum user involvement. Similarly, the imperative of transparency, which is essential to maintain a high level of trust in our information spaces, is limited by the desire of platforms to minimise the information shared with the public about the operation or effects of their platforms.

These limitations call for a legally enforceable regulatory framework that emphasises transparency and greater accountability on the part of online platforms and search engines. France is one of the few EU Member States to have pioneered this framework, making online platforms more accountable for the risks of manipulating information and spreading hate content on their services. As the body responsible for ensuring that these national laws are properly applied, Arcom has been able to build a rigorous dialogue with providers and develop expertise in regulating online platforms. The reviews it has carried out since 2020 of the resources put in place by providers show that the initial results are encouraging.

A collective framework, the DSA, will now apply to all digital intermediaries operating in Europe, increasing the responsibilities of very large online platforms and search engines at EU level and mobilising a network of regulators, working together to regulate these systemic players in terms of their size.

B. The DSA consolidates these shared achievements and establishes a collective framework for accountability and transparency.

The DSA is profoundly renewing the legal framework applicable to players in the digital economy.

It extends to all digital intermediaries the obligations that were previously expected only of very large platforms in terms of transparency, diligence in moderating illegal content and collaboration with third parties, while maintaining a proportionate approach, taking account of differences in nature and function between players within the digital ecosystem. For this new category of online platforms, the DSA introduces a system of obligations whereby the level of requirement varies according to company size (SMEs benefit from a simplified system), and which is particularly stringent for players generating systemic risks on account of the number of European citizens using them.

These new binding obligations include:

i) **the transparency and traceability** of orders to remove content or provide information issued by administrative and judicial authorities with regard to action against illegal content and the identification of alleged authors;

ii) **the strengthening of obligations in terms of the tools** available to the general public **for reporting** illegal content or content that contravenes the general terms and conditions;

iii) **the creation of the trusted flagger status**: the existing practice of using trusted third parties to identify and report illegal content will be governed by a specific statute recognising their special expertise and independence, and imposing stricter obligations on providers in terms of diligence in handling notifications sent to them by these third parties. This trust-based status will be accompanied by greater transparency in their activities, in particular through the publication of annual reports;

iv) **protecting the public (especially minors)**, by giving them the tools to participate, through their choices and actions, in making the digital space safer (information, platform settings, reporting, recourse) and in the use they make of it.

To deal with an operator who fails to comply with his obligations, the DSA provides for collective supervision of platforms in which all the regulators participate in a concerted action and, in the event of a proven breach, for the competent regulator (the European Commission or the competent authority of the country of establishment, as the case may be) to have a range of means of intervention, from the power of investigation to requesting the temporary restriction of access to the service from the judicial authority, via the imposition of financial penalties (e.g., fine and/or daily penalty payment).

C. For VLOPSEs, taking better account of systemic risks

The transition from a form of self-regulation that may be framed by initial regulations in certain Member States to a common regulatory framework overseen by the European Commission and mobilising a network of regulators in all the Member States is also warranted by the particular risks that very large online platforms and search engines are likely to generate because of their user numbers (over 45 million in one month), their uses, and the way they operate.

The scale and impact of these systemic risks differ from those attributable to smaller players. The multiplying effects specific to major networks potentially contribute to reinforcing the dissemination of illicit content, or their capacity to induce or amplify systemic risks is likely to cause lasting and serious harm to democratic values, weakening the quality of civic discourse, jeopardising public order, or unduly compromising the exercise of freedom of expression in these new public debate forums.

The identification, assessment and mitigation of these systemic risks, which are at the heart of the European regulation project, are reflected in the introduction of innovative instruments (risk-based approach and compliance, generalised transparency, initiated by the platform and third parties independent of it, notably involving auditors, the academic world and civil society), which are already in the pipeline in some national legislations but which are here consolidated and brought to European level.

Therefore, the DSA places the emphasis on risk audits by independent third-party auditors, the introduction of appropriate measures to mitigate these risks and the audit of such measures, access to data for experts in combating the dissemination of harmful content (in particular the academic world) and, in exceptional circumstances, the introduction of appropriate mechanisms to respond to crises, emergency measures to deal with extraordinary events involving a serious threat to the integrity of the Union or part of it.

D. Timetable and Arcom's place in the European architecture for regulating online platforms

This common framework will be developed and enriched over the long term, which is essential in terms of public freedoms.

In particular, it will be fuelled by exchanges between national regulators and the European regulator, as part of a collective dynamic supported by a rich dialogue with the European Commission and the Member States, and between national regulators and all stakeholders in each Member State.

This work has already begun: the designation of the first VLOPSEs in April 2023 by the European Commission marks a first step. The second will be the application of the above-mentioned provisions to these from 25 August 2023.

In France, the provisions inherited from the Act of 24 August 2021 will give way to the entry into force of the DSA applying to all digital intermediaries from the beginning of 2024.

This entry into force is being ushered in by the joint work of the competent authorities in France, as provided for in the Bill to Secure and Regulate the Digital Environment (CNIL, DGCCRF, Arcom), and by sustained exchanges with the Authority’s partners public authorities in combating hate contents such as the Interministerial delegation against racism, antisemitism, and hate against LGBTQIA (DILCRAH), Central Cybercrime Prevention Office’s PHAROS, the Sub-division in charge of Online Hate at the Paris Prosecutor’s Office, the secretariat of the National Consultative Commission on Human Right (CNCDH) and the National Digital Council (CNNum), civil society, the academic community and platforms.

Appendix 1

List of recommendations made by Arcom to combat the dissemination of hate content online

On transparency in general

- **Recommendation 1:** Strengthen the appropriate resources, in particular procedural, human and technological resources, in order to meet transparency obligations diligently.

On accessibility, transparency, and clarity of terms and conditions

- **Recommendation 2:** Make the pathway for users to access the general terms and conditions of service clearer, faster and more fluid.
- **Recommendation 3:** Improve legibility and navigation within and between the various pages devoted to general terms and conditions, Community rules, the help centre (or equivalent), etc.
- **Recommendation 4:** Clearly specify, in the general terms and conditions of service, which types of content and behaviour are prohibited by national law and the operator's rules, and in particular the prohibition of incitement to hatred and online harassment.
- **Recommendation 5:** Be as clear and explicit as possible, within the general terms and conditions, about the existence and operation of the internal mechanism for appealing moderation decisions, both for people who report content and for those whose content is moderated.

On the accessibility and intelligibility of reporting systems

- **Recommendation 6:** In transparency reports, please include:
  - hate content reported via *ad hoc* forms and via the content reporting system;
  - those removed (i) based on the general terms and conditions and (ii) under national law.
- **Recommendation 7:** Improve the accessibility of reporting systems, in particular by using more explicit symbols and headings.
- **Recommendation 8:** Illustrate the headings of the reasons for the report with concrete examples.
- **Recommendation 9:** Provide a brief reminder of the rules in force on the service in terms of content before users finalise their reports.
- **Recommendation 10:** Provide users with the option of attaching a description of any element they consider useful for analysing their report.
- **Recommendation 11:** Provide users with the option of indicating which part(s) of the content they wish to report.
- **Recommendation 12:** Allow users to report several items of content from the same account in a single report.

- **Recommendation 13:** Allow users reporting video content to indicate a complete time interval, not just the start time stamp.
- **Recommendation 14:** Provide users with the option of indicating whether their report also applies to links included in the description of the content reported.
- **Recommendation 15:** Be vigilant and, where necessary, maintain a capacity to react to misuse of reporting tools.

Furthermore, in keeping with what it has already indicated in its guidelines, Arcom reiterates the following recommendation:

- **Recommendation 16:** Allow users to indicate whether they wish to be kept informed of developments in the processing of their report.

On the means used for moderation

- **Recommendation 17:** Increase the transparency of moderation policies by making public the number, working language, and location of moderators employed by the operator.
- **Recommendation 18:** Ensure that human resources dedicated to moderating illegal content are adequately sized.
- **Recommendation 19:** Systematise, in transparency reports, the distinction between the origin of the report between users, trusted flaggers, and public authorities.
- **Recommendation 20:** Collect and make public data enabling the speed with which notifications sent by trusted flaggers are processed to be objectively assessed.
- **Recommendation 21:** Allow any user to challenge a moderation decision and ensure that these appeals are treated equally, without any particular legal argument being required.
- **Recommendation 22:** If the reversal rate, following an appeal, of moderation decisions or actions is high (as is the case for *TikTok* and *Dailymotion*), take appropriate measures to assess the relevance of the initial moderation and remedy it if necessary.
- **Recommendation 23:** Allocate a number of analysts commensurate with the need to process requests from public authorities promptly.

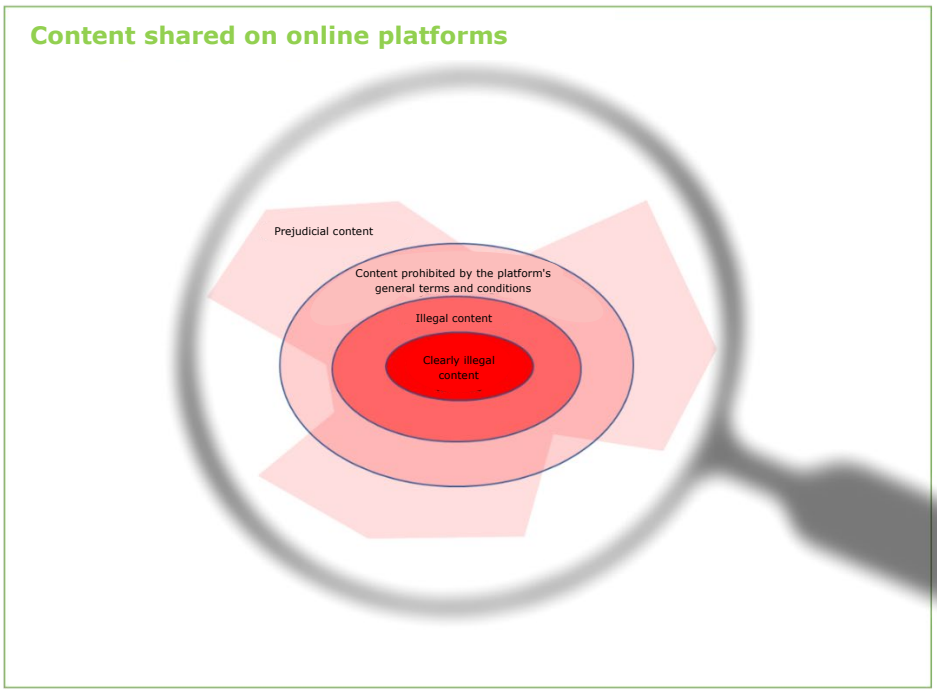
Appendix 2

Moderating illegal and harmful content on online platforms

Users of an online platform are required to share only content that is not prohibited either by law or by the platform's general terms and conditions (GTC), which are often more restrictive than the law.

When a platform becomes aware of content or behaviour that may contravene the law or its GTC, either as a result of a report or because it has detected it itself, it must examine it to determine whether it does in fact contravene the rules. The content or conduct in question may be manifestly unlawful or in breach of the GTC.

Certain content or behaviour may be aggressive, disturbing, or unpleasant for some or all of the users who see it, but this is not prohibited by law or the GTC. However, their proliferation or the increased visibility of the platform may contribute to inducing or magnifying a systemic risk. For example, the proliferation of aggressive content may contribute to a sense that digital spaces welcome bullying behaviour. Other examples: inauthentic behaviour may, on a large scale, weaken confidence in civic discourse and processes. The platform does not necessarily have to remove the content in question or prevent the user from expressing their views; however, some platforms opt to limit the virality or algorithmic propagation of this type of content, by removing it from their recommendations, for example.



Online platforms put in place various moderation actions which, depending on the case and the severity of the issues, may affect either the content or the account. The process implemented depends on each platform, which is free to decide the order and scale of interventions.

However, in the interests of respecting users' rights and creating a trusted digital environment, the DSA requires platforms to clearly and transparently explain this process to users, and to allow and facilitate the exercise of recourse in the event of disagreement with a moderation decision.

Actions on content	**Advance warning(s)**
	Temporary or permanent demonetisation of content
	Content maintained with warning next to or on the content (screen, filter)
	Reduced visibility: content dereferenced, removed from recommendations, visible only to "friends"
	Making invisible: content masked for all or removed
Actions on an account	**Advance warning(s)**
	Temporary or permanent suspension of access to the option to monetise account content
	Account and content dereferenced, removed from recommendations, visible only to "friends"
	Write block (account visible but rendered inactive)
	Suspension of the account for varying lengths of time or even permanent deletion of the account

Different types of moderation action possible *(non-exhaustive)*