COUNCIL FOR
MEDIA SERVICES

DSA TRANSPARENCY REPORTS

BRIEFING

31/01/2024

Jakub Rybnikár & Katarína Drevená

Introduction

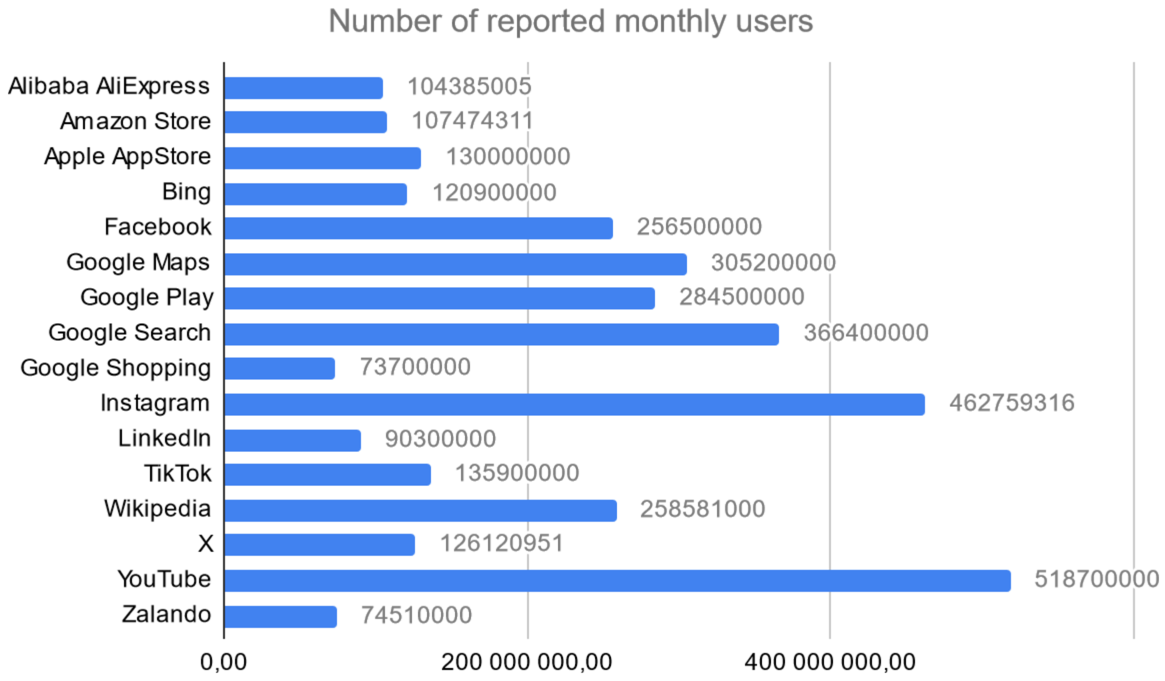
The Digital Services Act (DSA) is a crucial piece of EU digital legislation that tackles online harms and improves the standing of individual users vis-a-vis the largest online services. As part of the new regime, all intermediary providers (i.e. all services covered by the regulation) are obliged to publish annual transparency reports concerning their respective content moderation efforts. While this particular obligation kicks in only on the 17th of February 2024, platforms designated as VLOPSEs (Very Large Online Platforms and Search Engines) have to publish biannual transparency reports with additional information concerning primarily the human resources dedicated to content moderation in each of the EU member states. As article (art.) 42 DSA states, VLOPSEs must publish their first transparency reports at the latest by two months after the obligations referred to in the DSA had begun to apply.

For this, the first batch of VLOPSEs (19 services) published their first transparency reports in the last week of October 2023. Considering the speed at which the regulation has been implemented, it is no wonder that the published transparency reports lack any sort of standardisation regarding the structure or format, as well as a shared understanding of the required metrics. The reports comprise hundreds of pages filled with often incomprehensible qualitative and quantitative data. Having read through all of the reports, it is safe to presume that effective and actionable transparency cannot be achieved without a standardised format and simplified language so that a wide range of stakeholders can read and utilise the findings of these reports.

As the Slovak media regulator, the Council for Media Services (CMS) is responsible, as per art. 110 of the Slovak Media Services Act, to conduct analysis and research to map and apprehend the media landscape. Considering the novelty of the transparency reports, as well as the historic opaqueness of the VLOPSEs, CMS prepared a short brief on the contents of the transparency reports to help stakeholders navigate the current trends in content moderation and platform functioning.

This brief covers 18 out of 19 regulated services, with Pinterest being left out of the analysis due to the complexity and length of its report [1]. In an attempt to generalise findings and avoid comparison, the brief presents the most important data from the transparency reports and focuses on the services with the highest number of reported users. Considering the difficulty of exporting the quantitative data from the reports, we provide all the data in .csv format here to facilitate research and further analysis of the findings.

[1] Pinterest's report includes different vocabulary and metrics than the rest of the VLOPSEs. Additionally, it is not possible to easily extract the quantitative data from the report which makes any sort of analysis unfeasible



Main findings

Government orders

The first transparency reporting obligation under the new DSA regime, set out in art. 15 (para. 1a), is for all intermediary providers to report **the numbers of orders received from Member States' authorities**, including orders issued in accordance with Articles 9 and 10. The orders ought to be categorised by the type of illegal content concerned, the Member State issuing the order, the median time needed to inform the authority issuing the order, as well as the median time needed to give effect to the order.

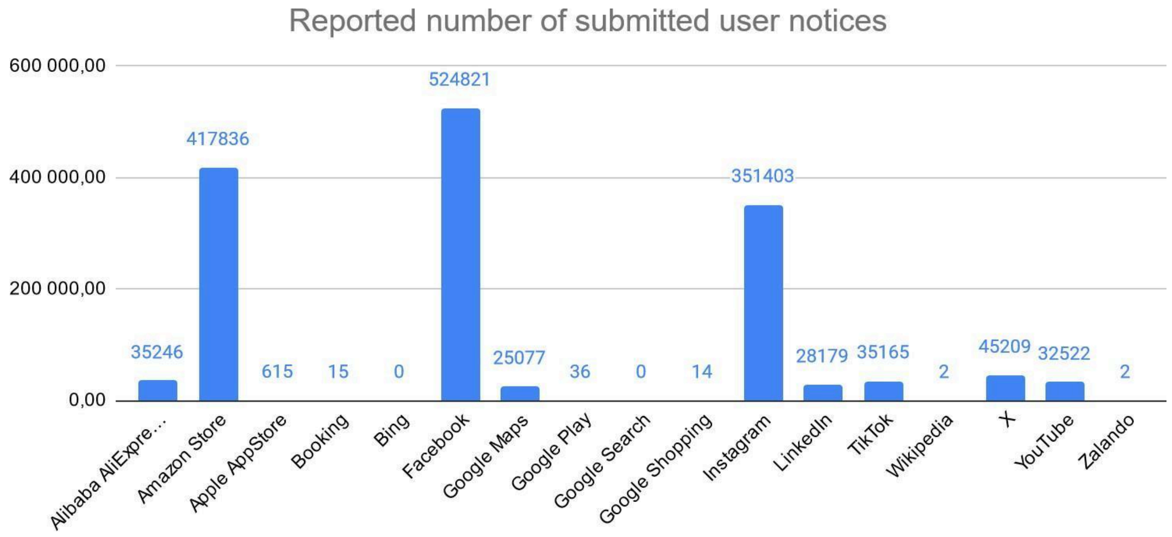
From the 18 analysed VLOPSEs, all but one (Snapchat) reported the desired information successfully. In practice, most of the analysed platforms issue an immediate confirmation receipt and only then proceed to analyse the order. The reports show that almost all platforms treat governmental orders as a top priority in their content moderation efforts and seek to resolve them as soon as possible.

For this, the median time needed to give effect to an order is generally below **seven days**. Not surprisingly, orders concerning product safety get resolved within two days, while orders concerning harassment require **seven days**.

What is, however, worrying is that all major social media platforms report only a few or no art. 9 orders. At the same time, however, the same companies report hundreds of art. 10 orders. This hints at a range of possible issues, including a lack of clarity concerning the reporting obligations, a fragmented set of definitions with each company treating these orders differently, or simply the inability of regulatory authorities to effectively process and issue content removal orders. Moving ahead with the implementation of the DSA, it will be interesting to monitor the flow of governmental content removal requests, especially in smaller markets where platform harms are reportedly most profound.

Notices submitted under art. 16 DSA

Considering the importance of user-flagging in content moderation, DSA requires hosting and platform providers to report the number of received user notices submitted in accordance with art. 16, as well as any action taken pursuant to the notices. A brief look at the transparency reports reveals that all platforms have received at least some notices lodged via a dedicated mechanism established by art. 16. While some platforms, especially social media platforms, received relatively large numbers of such notices (Facebook 524 821, Instagram 351 403), others had to barely process any (Booking 15, Zalando 828). In respect of the median time needed to process such notices, some VLOPs perform significantly better than others. Nevertheless, most VLOPs manage to process such notices within hours of receiving them (TikTok 13 hours; Instagram 20,4 hours; LinkedIn 23 minutes)



Based on the reported information, it seems that the effectiveness of the systems established pursuant to art. 16 DSA might not suffice for the intended goal. For one, most VLOPs actioned on a relatively small number of notices. For example, Apple Appstore actioned only on 2% and X on 19% of all submitted notices. This might have been, at least to some extent, caused by the poor quality of notices, but a brief glance at the format and structure of the notice mechanisms available via the individual platforms hints at a potential attempt to discourage users from submitting such notices in the first place. While undoubtedly subjective, this evidence is supported by our regulatory experience with the notice and action mechanisms available via the interface of several VLOPSEs. Second, mechanisms established under art. 16 should allow **any entity** to submit a notice. In the case of many analysed VLOPs, such as AliExpress, it is impossible to lodge a notice without an account. Third, the wording of the reporting obligation prevents VLOPSEs from reporting on the type of action taken with respect to the category of the alleged content. This prevents regulators from assessing the performance of VLOPSEs on a more granular level[2].

Additionally, the reports reveal a few interesting details that ought to be mentioned but are beyond the scope of this brief:

- The largest VLOPs, especially social media platforms, claim to process all notices manually. It is arguably difficult to comprehend how the small content moderation teams, reported as per obligation in art. 42(2a) DSA, manage to assess the notices adequately.
- In some instances, VLOPs do not report on the actions carried out pursuant to the notices but rather subsume them in their voluntary content moderation measures.
- Granularity of reported data seems to be rather problematic when dealing with multi-service providers (e.g. Google).

[2] [2]Albeit, it may be argued that this is partly solved by the SOR database.

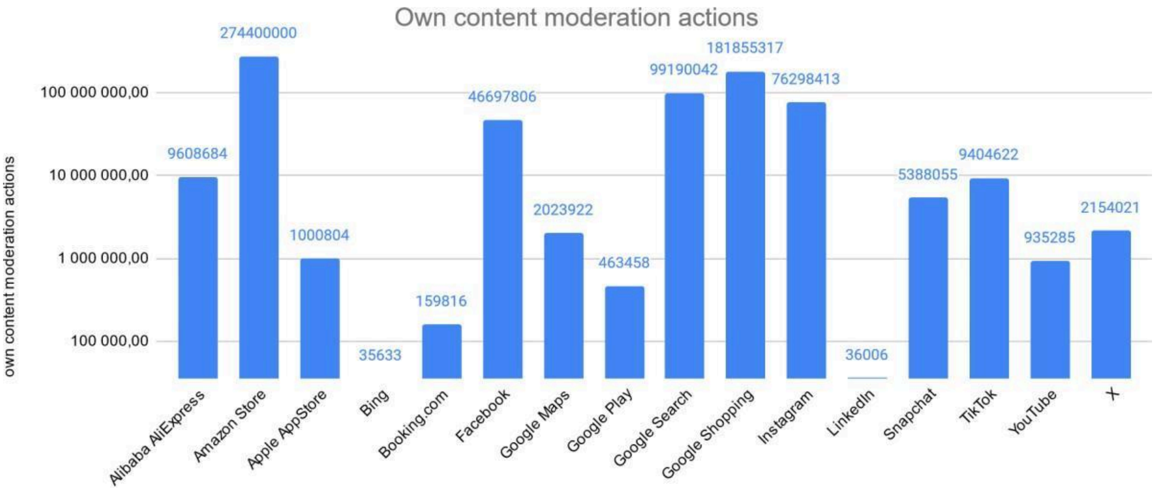
Own-initiative content moderation

Until recently, the regulated services published voluntary transparency reports based on the Santa Clara transparency principles that included very little comprehensive information on the procedural and technical aspects of content moderation/policies enforcement. The new transparency regime requires VLOPSEs to provide meaningful information about the providers' own content moderation efforts, including granular data on the number and type of measures taken that affect the availability, visibility and accessibility of information.

Without much surprise, the reports showcase a particular trend in content moderation that cuts through the VLOPSEs based on the type of service they provide. On one hand, marketplaces, such as Amazon or Aliexpress, focus heavily on scanning their respective services for any content that may violate their ToS or EU laws. From the data provided, it is clear that this pays off as scanning for prohibited products is fairly accurate and does not raise a number of fundamental rights issues. On the other hand, social media platforms refrain from any such scanning and instead focus on a wide range of detection methods, emphasising algorithms and external contractors that flag content. Nevertheless, most of the analysed VLOPSEs dedicate significant resources to conducting ex-ante checks while avoiding ex-post mitigation measures. It is important to note, however, that the information provided by the VLOPSEs regarding the detection methods is often too generic and available only in **55%** of all reports. Coupled with the lack of standardisation, this poses a substantial barrier to understanding the procedural and technical measures feeding content moderation practices.

With respect to the actual enforcement actions, all VLOPSEs prefer **limiting the visibility** of violative content rather than removing it. Considering the fact that this restriction is applied, perhaps bona fide, across the whole spectrum of categories of violative content, it may be argued that such an approach is problematic and may further exacerbate, for example, hateful conduct online. In other words, such content moderation actions limit the **spread** of violative content but do neither effectively punish nor preclude the **publication** of such content, which does not address problems related to filter bubbles and/or radicalization networks.

In terms of categories of content that get actioned on the most, the reports reveal a shared paradigm across all the services. Overall, the VLOPSEs target primarily hate speech, adult nudity and sex, as well as violence and incitement. However, looking exclusively at the number of enforcement actions the VLOPSEs carried out in the reporting period, marketplaces, namely Amazon and Google Shopping, outperform all other platforms by a significant margin.



Complaint handling

One of the key goals of the DSA is to foster parity between service providers and their recipients. To do so, the DSA codifies a complaint-handling mechanism that ought to be operated by all platforms and allow users to contest content moderation decisions. As part of the transparency reports, the VLOPSEs must report on the number of complaints received as well as on the decisions taken pursuant to these complaints. Additionally, the platforms must publish the basis for those complaints, the median time needed for taking decisions on them and the number of instances where those decisions were reversed.

Overall, almost all platforms report the required information (91% average compliance rate across all four reporting criteria). Insofar as the contents of the reports go, there can be found only very little qualitative information on the procedural functioning and efficiency of the complaint-handling mechanisms. Treating this metric exclusively quantitatively prevents regulators from making a qualified judgement about the efficacy of the mechanisms. Nevertheless, the available quantitative data indicate that a complaint has a fairly **high chance of overturning** the original decision once submitted.

Considering the sheer number of complaints the VLOPSEs receive each month, it is interesting to observe the median time needed for the VLOPSEs to take action. In general, the median time is around 24 hours, with LinkedIn having the shortest reaction time (**49 minutes**). Besides this, the basis of most complaints is a prior content moderation action taken by the platform rather than its alleged negligence of notices submitted via a mechanism established in art. 16. This is well demonstrated in a correlation between the high-volume categories of own-initiative enforcement actions and the categories of content warranting a review after a complaint.

Automated content moderation

The published transparency reports reveal an extensive reliance on automated means of content moderation. As a matter of fact, automation is responsible for over 90% of all content moderation actions taken by Meta's social media platforms (Facebook 93%; Instagram 98%). Other social media platforms utilise it to a lesser extent while enforcing their own Terms of Service (X 76%; TikTok 44%). Surprisingly, considering the reported scanning of content prior to publication, marketplaces resort to automated content moderation in roughly a quarter of all enforcement actions (Amazon 27%; Apple Appstore 22%).

Overall, all companies display an adequate compliance rate (average 84%), with the lowest (72%) being the description of **accuracy indicators for each applicable MS language**. When reported, the major online platforms all report high accuracy rates with stable intervals (90%+) in all applicable MS languages (LinkedIn is an outlier with 8-14,7% rate of error globally).

As in the previous part, the reports disclose swathes of data that warrant more attention than this brief allows. For this, we provide a bullet point list with interesting qualitative findings:

- All social media platforms utilise a combination of both automated and manual review, with automation being used for the most graphically discernible categories of violative content (e.g. child pornography or CSAM).
- Google's automated enforcement actions are rarely overturned once contested.
- Social media platforms, such as Youtube and Meta's Facebook and Instagram use hashing to prevent the re-upload of violative content (i.e. flagging or removal of content that is similar to content deemed violative).

- TikTok manually reviews all pieces of content once they reach a particular virality.

Suspensions

According to art. 24 (1b) DSA, providers of online platforms shall provide information on the number of suspensions enacted for the provision of manifestly illegal content, as well as the number of suspensions for the submission of manifestly unfounded notices and manifestly unfounded complaints.

Two out of nineteen VLOPSEs did not provide the required information regarding measures and protection against misuse of online platforms (Bing and Alibaba AliExpress), while the rest of the VLOPSEs provided information only on some of the requirements. Overall, platforms have similar policies in this regard, although some differences can be observed. For example, some VLOPSEs (e.g. LinkedIn) may permanently suspend an account after a single serious content policy violation; others (e.g. Instagram) do it only when their ToS are violated frequently (although not stating what ‘frequently‘ means). Similarly, some VLOPSEs (e.g. Facebook) may suspend for a limited period of time those users who repeatedly post manifestly illegal content, while others (e.g. Apple AppStore) prefer to terminate such accounts.

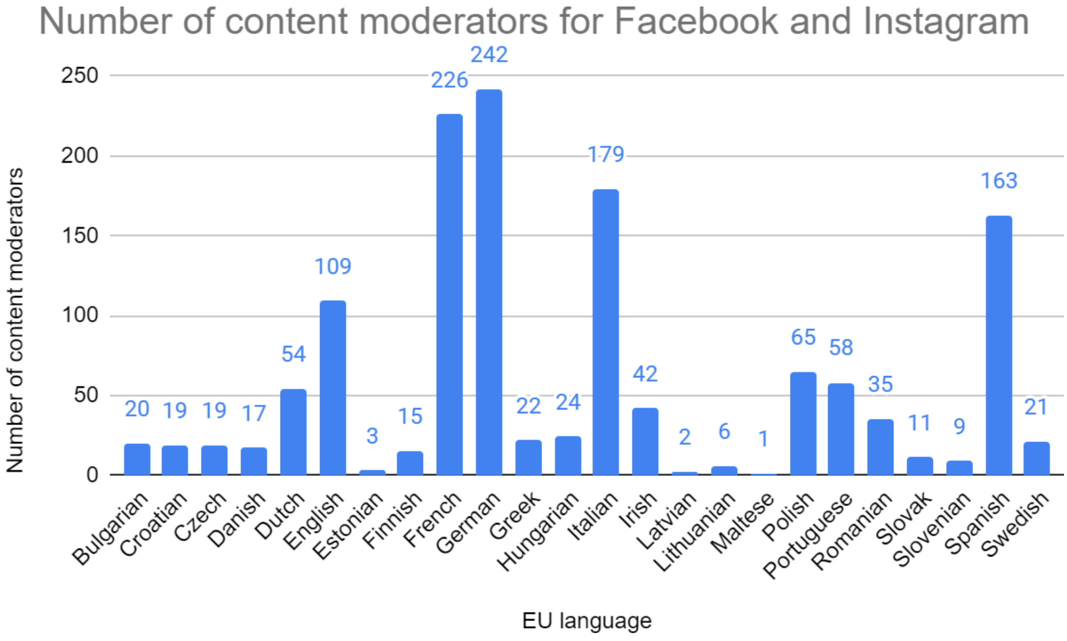
Based on the transparency reports, it is clear that several VLOPSEs did not suspend any user for either manifestly illegal content or submission of manifestly unfounded notices or complaints. On the other hand, a few VLOPSEs did so in hundreds or thousands of cases. Although the numbers are not comparable, it raises the question of whether appropriate suspension mechanisms are set up on all platforms so as to prevent the re-upload of violative and illegal content.

Human resources

Content moderation is considered to be crucial in ensuring a safe and secure online environment. Based on the transparency reports, social media platforms rely, to a significant extent, on automated means of content moderation, which makes any information concerning human resources all the more valuable. VLOPSEs are required to specify the number, qualifications, and linguistic expertise of persons responsible for content moderation, as well as the training and support given to such staff. Additionally, VLOPSEs shall provide indicators of content moderation accuracy.

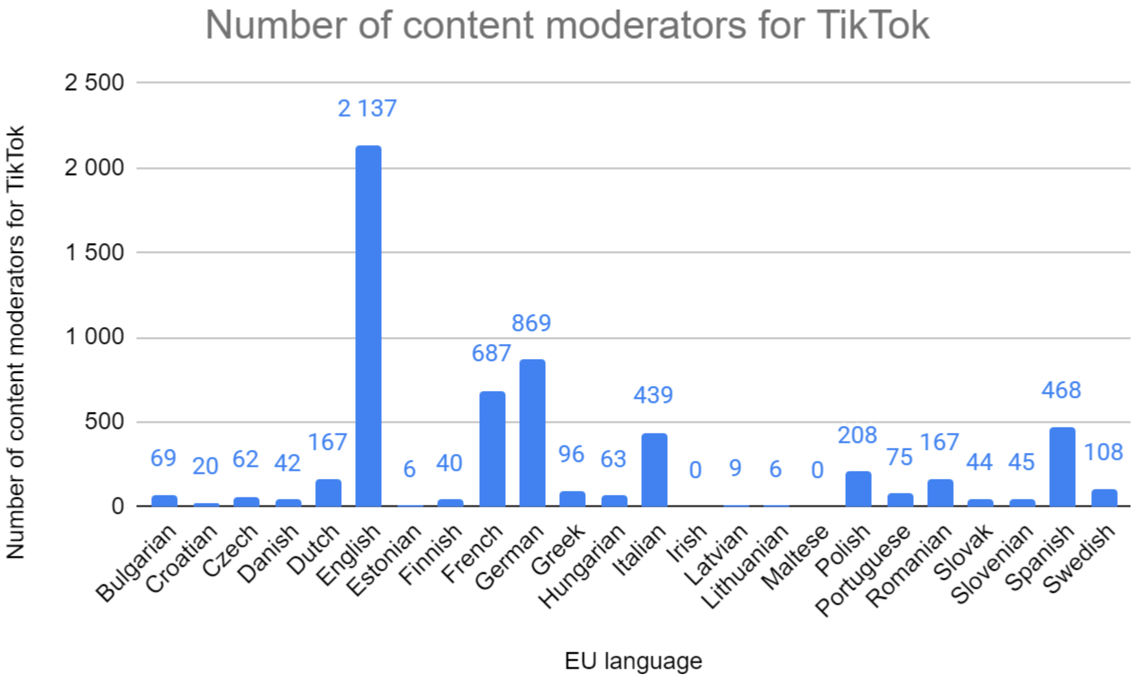
Overall, the majority of VLOPSEs provide regular training to their human reviewers to ensure correct assessment. According to the reports, the moderators are required to attain certain skills, education, and language proficiency, but the answers are vague, so no meaningful transparency can be derived from the answers provided by the VLOPSEs. Similarly, most VLOPSEs claim that people responsible for content moderation receive support, such as physical well-being activities or psychological support.

As for the number of content moderators, some VLOPSEs did not provide any concrete information (e.g. Bing) while some did not give data broken down by official languages of the EU MS (e.g. Alibaba AliExpress). Google services (except Google Maps), Facebook, Instagram and TikTok cover all EU MS languages, but they all seem to lack content moderators proficient in Irish and Maltese. The remaining VLOPSEs focus only on a few major EU languages while neglecting the rest, which may hint at a potential failure of the so-called duty to care. From a regulatory perspective, this may warrant further investigation, especially on social media platforms like X, as some parts of the EU population may not be afforded the necessary online safeguards.



Additionally, the reports show that reporting just a number of human resources dedicated to content moderation might not be a comprehensive tell-all metric. For example, Apple AppStore or Google provided the number of moderators broken down by official language; however, this data does not represent the number of content moderators hired to review that specific language. In fact, content moderators can be assigned to review content in several languages (i.e. TikTok reports to have 44 moderators for Czech, Slovak and Slovenian). For this, it might be worth considering adding more granularity to the mix by, for example, providing FTEs for each applicable language.

Finally, even in the case of VLOPSEs that cover all EU MS languages in content moderation, there ought to be a discussion of whether the dedicated human resources are adequate. For example, bearing in mind that numbers are not comparable due to a lack of standardisation, Facebook and Instagram, each with more than 255 million users in the EU, have together 1 362 content moderators, while TikTok, with more than 135 million users, has 5 827 people responsible for content moderation.



Conclusions

This brief on the contents of the DSA transparency reports sheds some light on the key facets of the VLOPSEs' content moderation practices. While responsive to government orders, the discrepancies in receiving and resolving art. 9 orders raise concerns about regulatory clarity. The prevalence of art. 10 orders prompts scrutiny, especially in smaller markets, pointing to potential challenges in processing and issuing art. 9 orders.

Art. 16 notices highlight diverse user-flagging engagement, with social media platforms processing large volumes promptly. However, challenges lie in notice submission mechanisms and reporting granularity. The section on own-initiative content moderation reveals distinct approaches among VLOPSEs based on their service type, emphasising the need for standardisation and detailed reporting.

Complaint handling mechanisms, while widely reported, lack qualitative insights, raising questions about their efficacy. The role of automated content moderation, vital for platforms, exhibits varied reliance and error rates, necessitating a closer look at their effectiveness. Suspension policies among VLOPSEs show differences and a lack of transparency, which could potentially warrant further investigation. The importance of human resources in content moderation is underscored, emphasising the need for detailed reporting on staff qualifications, training, and support. In conclusion, while transparency reports offer valuable insights, ongoing refinement and standardised reporting are crucial to ensure the DSA's goal of fostering parity between service providers and users in the evolving digital landscape.