CDT Europe Contribution to European Commission Public Consultation: Draft Implementing Regulation laying down templates concerning the transparency reporting obligations under the Digital Services Act

The Centre for Democracy & Technology, Europe (CDT Europe) welcomes the opportunity to provide feedback on the draft Implementing Act and supplementary Annexes which lay down the mandatory templates for the transparency reporting obligations of providers of intermediary services and of providers of online platforms within the framework of the Digital Services Act.

As explored in detail in CDT's Making Transparency Meaningful Report, transparency reports can provide a better understanding of the overall environment and participation on a particular service, and allow users and others to see how different content moderation practices and government demands for content restriction have changed over time. Increased accountability and transparency are all the more pertinent, as recent trends indicate a decline in public facing transparency and companies are making it particularly hard for researchers to study and better understand the societal impacts of intermediary services and online platforms. For example, TikTok has recently limited the utility of several of its transparency tools; X (formerly Twitter) has removed access to API for researchers through disproportionate financial restrictions; and Meta's new API system lacks functionality, all of which speak to these trends.

In order to achieve meaningful transparency with the reports mandated by the DSA, it is essential that the reporting templates form part of a comprehensive transparency framework. This includes ensuring the data gathered in the report complements the research developed within the scope of Article 40, the outcomes of the risk assessments under Article 34, alongside the results from the independent audits foreseen in Article 37. It is also important to incorporate lessons learned from similar initiatives aimed at gathering data from intermediaries and online platforms in the aim of increasing transparency. For example, evaluation reports related to the Code of Conduct on countering illegal hate speech[1] have proved to be very limited in providing meaningful insight into how allegedly illegal hate speech is addressed despite the extensive quantitative data provided.

Though the draft implementing regulation and accompanying annexes are extensive, CDT Europe takes this consultative opportunity to outline our reflections on aspects that we believe would benefit from additional clarification and guidance.

Impact of the transparency reports

As a first point, CDT Europe reiterates that given the complexity of transparency and the sheer scale of data collection that the DSA requires, there will be a need for a continuously iterative process in the development of the related reporting templates. CDT Europe strongly encourages the European Commission to reflect on how the requirements laid down in these reporting templates will impact on how reporting adapts in the future.

[1] https://www.euractiv.com/section/digital/news/progress-stalls-on-commissions-online-hate-speech-efforts/

For example, the issue of how data collection is weighted (e.g. significantly more emphasis on quantitative over qualitative) will be taken into consideration by intermediaries and online platforms, and this will be reflected in the quality of data received over time. For instance, persistent emphasis on the disclosure of key metrics on the rate of removal or response time of content will incentivize online service providers to focus on this metric - potentially leading to over-removal of lawful speech- despite not being a legal requirement of the DSA. Existing research[2] has shown that the unit of measurement for reporting on median time of response to orders for takedown for illegal content isn't always helpful because speed doesn't seem like the most helpful gauge of efficacy or accuracy. This is one aspect, but overall the standards established with these reporting templates will significantly influence future reporting trends, therefore we would also encourage more opportunities for future consultations on the usefulness of the data gathered through these reporting templates, given the iterative nature of the DSA.

Importance of Meaningful Transparency

CDT Europe welcomes the inclusion of quantitative and qualitative data in the reporting templates as both elements are essential to developing a more in depth understanding of how user-generated content is being moderated and more insights into broader content moderation practices across service providers. We also welcome the emphasis placed on establishing a transition period which will swiftly bring reporting periods into alignment early into the lifecycle of the DSA which will be crucial for coherency between enforcement agencies and in the development of research. However, in order to ensure these efforts result in meaningful data collection and transparency, clarification is required in the structure, categorisation and data points requested of the reporting templates.

As a first reflection, there is a need for clarification on Section 1.2.1 of Annex I and Section 1 of Annex II which outlines the categories of illegal content applicable to all sub-sections of the Quantitative Template. Though categorisation will be useful in compiling the reports, the categorisation of alleged illegal content and content in violation of a platform's terms of service is not clearly delineated. Specifically in Section 1.2.1. Annex I includes a column entitled "*Description of other*", which we deduce is reserved to identify content that is lawful but considered harmful or in violation of a platform's own content policies, however this title does not make that clear. Similarly Section 1 of Annex II includes 3 columns for "*category label*", "*category description*" and "*category of illegal content*", logically therefore, given that not all data requested in these reports relate to allegedly illegal content, there should be another distinct column or method to clearly categorise '*incompatible content*". Making this distinction explicit is important because it will increase the accuracy of companies' reports under the DSA and will give an idea of how often content is removed due to legal obligation rather than any other reason.

Alongside this practical observation it would be beneficial for the European Commission to provide explanatory guidance for how certain aspects of the reporting template should be

[2] https://www.cigionline.org/articles/how-transparency-reporting-could-incentivize-irresponsible-content-moderation/

interpreted. A primary example is to provide insight into how the proposed categorisations were determined and how they should be understood through, for instance, a legal reference point, definition or examples. Another more granular example can be found in section 1.2.1 which outlines that providers of intermediary services and online platforms should both "*indicate the number of items that have been granted or complied with*' as well as "*moderated based on orders received from Member States*'. It would be useful to understand what is being understood as 'compliance' and 'moderated' and why they have been differentiated in this context. There could be circumstances where a service provider may not act on an order due to the requirements of Article 9.2 (a) not being met, however may choose to 'moderate' said content on the basis of ToS violations. At present, such scenarios are not captured within the reporting template and it would be useful to understand when such circumstances arise.

Establish a Stronger Alignment Between Quantitative and Qualitative Data

CDT Europe notes that the weighting between the quantitative and qualitative reports differ significantly, with more emphasis being placed on the collection of quantitative data, and even with this emphasis, there are several aspects of the quantitative reports that would benefit from additional key indicators or columns. Alongside this, the information requested in the quantitative reports are not fully supported by the qualitative template nor is there scope to provide additional qualitative explanations for the quantitative data requested, which will be necessary to understand the data provided.

It is most notable that the quantitative data being requested in relation to Article 9 and 10 orders require the least amount of granular information out of all the sections proposed, yet there is a clear need for much information in the public interest. For example, there is no scope to ascertain which member state authorities have issued the related orders, or scope to ascertain the number of orders to act against illegal content not granted/complied with, and qualitative reasoning for why. It would be important to collect information on instances when intermediaries and online platforms don't comply with a member notice of illegal content, or similarly where such orders fail to meet the safeguards and standards outlined in Article 9, for the purposes of public scrutiny and enforcement oversight.

Similarly, additional indicators or columns would also be useful in Section 1.2.2 on data gathered in relation to Member State Orders for information. Though it is evident several of the columns will be the same as Section 1.2.1, it would be beneficial to not only capture the alleged illegal content in question, but the reasons a government may ask for information on a recipient of a service, or how these orders are to be categorised based on the information provided in the statement of reasons associated to these orders. For instance, it would certainly be in the public interest if a Member State ordered a platform to provide information about people who attended a protest and under what categorisations these orders are being submitted particularly during times of civil unrest.

Facilitating the collection of disaggregated data

Another key reflection is that certain aspects of the reports should be modified to facilitate disaggregated data collection, in order for real transparency and oversight to be achieved. For example, alongside the lack of information on specific authorities issuing Article 9 and 10 orders, there is no opportunity to disaggregate additional information, quantitative or qualitative, on the data related to number of notices received from Trusted Flaggers and number of items moderated based on these specific notices. The Trusted Flagger provision, though important, still raises certain concerns due to the ability for law enforcement agencies to be designated as such whilst simultaneously being empowered to issue Article 9 and 10 orders. From a rule of law and fundamental rights perspective, it would be important to determine how often law enforcement agencies are submitting notices as Trusted Flaggers versus issuing Article 9 and 10 orders. Taking into consideration that said Trusted Flagger notices may be internally weighted as being as significant as Article 9/10 orders by companies, it is essential that disaggregated data is collected on these specific notices. This additional information on which entities submit Trusted Flagger notices will also provide important insight into the types of entities that have been designated with such status and how their notices are being prioritised.

One method through which this could be addressed is to replicate the level of granularity being requested in section 1.4 into other sections of the reporting template, or to at least provide scope for this additional information to accompany these prior sections. For example, it would be hugely beneficial to understand which content moderation actions were implemented in cases related to notices submitted under the Notice and Action Mechanism. Facilitating the collection of disaggregated data alongside creating coherency between the various databases and reports that will be gathered as part of the wider transparency obligations under the DSA will allow for the development of essential research, as well as contribute to more efficient enforcement of the Regulation.

Ensuring coherency of definitions & terms

Several of the reflections noted above also relate to a lack of definitional consistency across the draft implementing regulation and accompanying annexes. This extends to the terminology used in the explanatory guidelines in Annex II as well as the use of terms that are, appropriately, defined differently across industry. The draft implementation regulation would benefit from definitional clarifications from the European Commission or reference to how these terms are to be broadly understood.

The most notable examples of this is in the use of *'automated means'* and '*content moderation'* as terms that are used widely, but that in practical terms captures an enormous array of potential actions. Online platforms vary widely in terms of operation and scale, and without supplementary information on how platforms are defining these elements internally, it will be difficult to facilitate comparative analysis of the transparency reports or to properly ascertain compliance. For example, what does "*number of items moderated*" mean for online platforms such Wikipedia or

Reddit, where more users are empowered to "*moderate*" or where a company may include a range of actions that are distinct to the platform under the moderation umbrella? Similarly, a wide variety across platforms in determining '*error rate*' is to be expected, and without scope to provide supplementary qualitative information, it may prove impossible to determine the accuracy of the 'error rate' data that has been provided. For example automated errors can include instances in which content is erroneously flagged as illegal content, as well as content flagged through the notice and action mechanism and subsequently restricted or 'moderated' by an automated-decision system but was later restored after human review.

Another notable example is the use of '*soft moderation restrictions*' in the qualitative template. Given the breadth and diversity of service providers that will be required to provide information for these qualitative reports, it's pertinent to note that a content moderation action that would be considered '*soft*' on one platform may not be considered as such on another. Similarly, a more detailed explanation of what is understood as a "*specific and objective reason"* for potential omissions from the reporting templates, as indicated in Part 1 Section 1 of Annex II, would be useful for all stakeholders. It would seem that in both cases the European Commission has an internal perspective or understanding of these terms and it would be useful to provide more insight in this regard.

To reiterate, it is clear that consistency across industry of these definitions is not possible given the diversity in the functionality of various platforms, however it would be useful for reporting platforms to be given scope to provide directly or link to information on how they internally define and understand these terms; this would be particularly useful for researchers who can reference these additional resources. It may also prevent the real risk of the development of studies that purport to reach definitive conclusions about content moderation across platforms that really won't be supported by the data available or creating false equivocation between companies' reports. What would it mean if TikTok restricted visibility of 200 instances of "*online bullying/intimidation*" and Meta similarly restricted 500 instances if those two platforms don't have the same definition for that category of content?

Overall therefore, there are certain terms used in the implementing regulation and accompanying annexes that need to be clarified by the European Commission. On the other hand, intermediary services and online platforms will need to supplement the data provided with sufficient qualitative explanation and information on how they internally understand certain terms.

Conclusion

CDT Europe understands the complexity in trying to capture all of this data and information in one reporting template and advocates to ensure the transparency reporting process is both practical and meaningful. Therefore, the above recommendations are not aimed at creating a burdensome reporting process, but to ensure the reports capture the most pertinent data points. This must be accompanied by greater alignment with information gathered in the context of other

transparency obligations under the DSA. For example, the information gathered in section 1.5.1 such as the data on decisions reversed, upheld and omitted in the out of court dispute settlement bodies should in some capacity link to the reports that will be developed on the functioning of these bodies. This extends also to connecting these reports with the existing database for the Article 17 Statement of Reasons, the publication of other related transparency reports such as those from auditors, as well as the other public interfaces being established by the European Commission and national regulatory bodies. To reiterate, it is essential that these transparency reports do not function as standalone data points, but be supplementary to all other relevant transparency provisions.

Our recommendations reflect what we discern as the key areas within the draft Implementing Regulation that should be strengthened in order to make sure that all the transparency reporting provisions are effectively implemented and result in the collection of tangible data that meaningfully creates increased transparency and accountability.