Study on transparency reporting of online intermediary services*

Ben Wagner, Marie-Therese Sekwenz, Johanne Kübler and Carolina Ferro

September 2023

* This study was drafted with funding from the European Commission. It is published by Enabling Digital Rights B.V (https://enabling-digital.eu/).

Table of Contents

1. INTRODUCTION AND OVERVIEW

This Study on transparency reporting of online intermediary services provides a comparative analysis of different forms of transparency reporting of online platforms. Using the EU Digital Services Act (DSA) as a starting point, we look at different transparency reporting requirements and practices by online platforms (DSA) (*Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and Amending Directive 2000/31/EC (Digital Services Act) (Text with EEA Relevance)* 2022). In order to do this effectively, we first develop a theoretical framework of the quality of transparency reporting based on academic literature. We then apply these quality categories to the transparent reporting requirements of online platforms, to assess which of the requirements and practices are likely to produce high quality outputs.

In order to conduct this comparative analysis, we have developed a set of criteria to target quality transparency reporting to assist in our analysis. When phased using these developed concepts, we conduct a comparative analysis of the different transparency-norm classes mapped against the quality categories for reporting, indicating reporting times per quality category. This comparative analysis is used to indicate the quality levels that map the transparency-norm classes per quality category, which specifies benefits and drawbacks per quality category. Additionally, we propose the use of an overarching concept for moderation basis classification: the content moderation source. The content moderation source should make reporting classes across services available that go in line with Art 17 DSA's Statement of Reason (SOR).

Transparency reporting is a leading concept for the governance of online platforms (MacCarthy 2020a). The DSA includes provisions to harmonise such reporting obligations within the European Union. Of special importance to the transparency reporting obligations of the law are the mandatory reporting duties as defined in Art 15, Art 24, and Art 42 DSA, which describe the extent of the reporting according to the platform in correlation of their service provided and users acquired. Their reporting depth, therefore, varies and is becoming more rigorous for online platforms and most excessive for large online platforms and search engines (Dinar and Hinrichs 2022). For example, Art 42 DSA would define reporting obligations that only are addressed to Very Large Platforms (VLOPs) and Very Large Search Engines (VLOSEs). Both VLOPs and VLOSEs, are defined in Art 33 para 1 DSA and take the number of average monthly users into account to measure the threshold of VLOPs and VLOSEs. Additionally, platforms can implement voluntary transparency reporting, e.g., according to Art 44 DSA (including Standards) or follow certain Codes of Conduct as described under Art 45, 46, and 47 DSA.

This study includes obligatory and voluntary reporting (see transparency-norm classes 1-4) and maps it against key concepts of the transparency reporting processes across the legal

landscape. By providing a novel analytical framework for evaluating the quality of transparency reporting, this study aims to provide guidance for stakeholders, like regulators and the public, to assess the reporting quality of platforms. Furthermore, by comparing and mapping transparency reporting this analysis will be performed on a set of selected norms and codes and provide categories to assess key concepts to map against within in the process of transparency reporting. This comparison includes the creation of best practices and recommendations for qualitative transparency reporting and proposes a visualisation for providing such information in a tabular form (see Annex). Additionally, the benefits and drawbacks of reporting are defining a selection of the best practices included in this report.

Outline and Overview of the Report

This section provides an outline of this report's structure, the methodology, the introduced definitions, and categories, followed by a comparison of existing transparency framework categories, recommendations, and best practices of transparency reporting, including their challenges and benefits.

The Report, therefore, uses coherent metrics across legal norms that are key indicators for transparency reporting to base its recommendations upon and conclusions on the information gained by following the structure to be presented in the next section.

1.1. Methodology

This section provides the proposed methodology. The aim is to create a novel analytical framework for evaluating the quality of transparency reporting. Therefore, we propose a new approach to map the quality of transparency reporting systematically by comparing and visualizing (in transparency reporting tables) quality level across transparency-norm classes.

The transparency-norm classes are:

2. The Digital Services Act (*Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and Amending Directive 2000/31/EC (Digital Services Act) (Text with EEA Relevance)* 2022);
3. Regulation (EU) 2021/784 on Terrorist Content Online (*Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on Addressing the Dissemination of Terrorist Content Online (Text with EEA Relevance)* 2021);
4. The German Network Enforcement Act (NetzDG) (*Gesetz Zur Verbesserung Der Rechtsdurchsetzung in Sozialen Netzwerken – Network Enforcement Act* 2017a);
5. OECD Terrorist and Violent Extremist Content (TVEC) Voluntary Transparency Framework (OECD 2021);
6. The Santa Clara Principles on Transparency and Accountability Around Content Moderation, and associated implementation toolkits for advocates, companies ('Santa Clara Principles on Transparency and Accountability in Content Moderation' n.d.);

7. Relevant codes of conduct with transparency measures, such as the Code of Practice on Disinformation ('European Economic Area (EEA) - Code of Practice on Disinformation' 2023);

The transparency-norm classes furthermore can be distinguished into four classes that indicate their reporting source:

- Obligatory Legal transparency-norm classes (1.),
- Voluntary Reporting transparency-norm classes (2.),
- Code of Conduct transparency-norm classes (3.),
- Voluntary VLOP/VLSE transparency-norm classes (4.)

These class-details are indicated in the proposed comparison of transparency framework categories in section 2. By providing a systematic and analytical framework to map the different demands and obligations identified within the set of transparency-norm classes that include legal sources (including EU law and Member State law), regulating initiatives and codes of conduct, these ten transparency-norm classes provide a good norm base for the analysing them under the quality category.

To set out the metrics under which the transparency-norm classes we compared within this report the transparency reporting qualities categories. These quality categories create a taxonomy of categories for the analytical framework. These key concepts are drawn from the literature this report is based on that express important concepts supporting quality of transparency reporting. Additionally, we could draw conclusions from previous work that influenced the creation of the categories. The quality categories were developed based on findings from prior research especially addressing contextualisation and comparability (Wagner et al. 2020a; Kübler et al. 2023) and our initial literature search for this report resulting in five quality categories:

1. Contextualization
2. Comparability
3. Vulnerability to Manipulation
4. Provability
5. Machine-Readability and Processability

The quality categories aim to summarize important concepts of transparency reporting across legal scopes, industries, and disciplines to create meaningful elements for the comparative analysis across the transparency-norm classes. The quality categories, therefore, will be described in detail in the following section focusing on each of the five.

By combining the transparency-norm classes with the quality categories the quality levels will be evaluated by using the insight gained per transparency-norm class that indicates the quality category including their benefits and drawbacks. The quality level, therefore, proposes an initial design for reporting tables that map out the transparency-norm classes against the quality category in a visual form supporting the comparative analysis. By mapping the

proposed methodology as quality levels this report aims to create a comprehensive analytical framework for evaluating the quality of transparency reporting.

This research methodology can by its comparative element also be understood as creating a Red-Team-testing approach for providing a new analytical framework for evaluating the quality of transparency reporting processes taking together qualitative and quantitative elements.

The quality levels therefore provide guidance on the form, the details and the content needed for evaluating the quality of transparency reporting. Furthermore, reporting times are defined and included in the quality levels to provide guidance and recommendations on meaningful reporting time frames. By including recommendations and best practices for transparency reporting this study defines a transparency reporting standard by considering the benefits and challenges certain options might bring to indicate the quality level per quality category and transparency-norm class in a tabular visualisation.

Furthermore, the report proposes a structure that considers the different reporting levels for online platforms regarding their service and size. The different reporting levels are referred to as reporting category. The reporting categories include five classes:

1. Intermediary Services
2. Hosting Providers
3. Online Platforms
4. Online Platforms with Means for Distance Contracts
5. VLOP/VLSE

By considering their varying levels of reporting obligations and scrutiny this report purposes an initial structure that considers the quality level per reporting category and allows, e.g. detailed information on their reporting times per category and quality level-detail.

Furthermore, the content moderation sources act as an overarching list of transparency reporting reasons, that support the reporting structure of Art 17 DSA and follow the reporting in line with the definition of SOR for supporting coherent and `seamless´ reporting across different reporting obligations. By linking transparency reporting structures to the SOR categories we support structured and efficient reporting that strengthens efficiency and comparability for increasing meaningful insight between compliance of transparency reports and SOR obligations of platforms. These content moderation sources developed for this report, therefore, indicate an initial list of reasons for content moderation action according to the SOR and the terms and conditions according to Art 14 DSA. The SOR should create an overarching reporting scheme that works across Member States and their diverging legal landscapes to make reporting better comparable. However, the DSA itself demands also to indicate information on `illegal content´ as defined in Art 3 lit h DSA. The legal reason, or specific national Member State norm, is another reporting detail in line with Art 17 DSA that

ideally can be added within reported information additionally to the content moderation source.

This set of content moderation sources should help to find comparable categories for transparency reporting across platforms that cover contractual categories and legal norms or norm classes. This set of standardized overarching moderation sources can help to increase comparability across platforms and across Member States on a higher level for transparency reports.

Content Moderation Source

This section provides a list of content moderation sources that aim to increase comparability across platforms and provides a set of legal sources for platforms to cover the main reporting categories for 'illegal content' and violations of overarching terms and conditions.

In order to take moderation decisions online platforms can base their decisions on sources of moderation: either in the form of contractual clauses (terms and conditions) or because of the law (Wagner et al. 2021). By analysing content moderation decisions of a small/medium sized platform we legally analysed a sample of content in comparison to the moderators behaviour and moderation reason. We also propose to include such content analysis techniques for empirically evaluating the quality of content moderation especially in line with systemic risk assessments according to Art 34 DSA (Kübler et al. 2023).
For the first time, the Digital Services Act (DSA) will harmonize content moderation efforts in Europe and the reporting of content moderation decisions according to Art 17 and the indication of SORs. Due to the definition of `illegality´ within the DSA in Art 3 lit h DSA legal sources stemming from different Member States might be a reason for moderation and indicated at a higher level in the overarching SOR categories and as here proposed under the content moderation source. The additionally demanded legal reference for Art 17 DSA will be referred to by two-letter abbreviations specified per content decision to indicate the Member State legal source ('Api Documentation - DSA Transparency Database' n.d.).

Additionally, it should be mentioned the reporting duty of VLOPs according to Art 31 para 1 DSA, and their systemic risks described in Art 34 para 1 lit a-d DSA. Referring to `illegal content´ (lit a), the risk to fundamental rights (lit b), the risk on 'civic discourse and [the] electoral process as well as the public security (lit c) and the risk that 'gender-based violence' poses, public health physical and mental well-being, or the protection of minors (lit d). It, therefore, would be advisable to also include the risk reporting categories in the indication of the SOR or in content moderation sources as proposed within this report. Additionally, the structure of indicating the decided upon content moderation action, e.g. negative effects on visibility or the monetisation of content referred to in Art 20 DSA that indicates details about the internal complaint handling system should be considered for the creation of SOR categories and content moderation sources for transparency reports. Furthermore, it should be mentioned that also the Notice and Action Mechanism as referred to in Art 16 DSA is an important part

of ensuring meaningful reporting and transparency that understands that the design of the flagging process is an essential part for the numbers indicated in transparency reports (Wagner et al. 2020a).

Because of the increased complexity of the potentially applicable Member State norms (Radu et al. 2021; Golia, Kettemann, and Kunz 2021; Amelie Heldt 2019), the SOR furthermore defines overarching categories or as indicated in this report, content moderation sources ('Api Documentation - DSA Transparency Database' n.d.). These content moderation source class includes 14 categories overall, incorporating violations against the terms and conditions according to Art 3 lit u DSA and Art 14 DSA, and an uncategorised category to collect reasons that otherwise might not fall within a definition of another category to increase flexibility to the development of events and moderation needs.

Category

This section provides an overview of the currently proposed SOR categories and moderation sources in this report to support transparency reports with an overarching structure in line with Art 17 DSA. This list of content moderation sources is based on the proposed categories selected for Art 17 DSA. It should be mentioned however that reporting also might consider the granularity of content moderation source per Member State and language. The categories include:

1. Animal welfare
2. Data protection and privacy violations
3. Illegal or harmful speech
4. Intellectual property infringements
5. Negative effects on civic discourse or elections
6. Non-consensual behaviour
7. Pornography or sexualized content
8. Protection of minors
9. Risk for public security
10. Scams and/or fraud
11. Self-harm
12. Scope of platform service
13. Unsafe and/or illegal products
14. Violence

The content moderation source referring to the terms and conditions which online platforms craft to contractually govern the different forms of speech that are uploaded online complete the legal reasons moderation can be based on. These reasons for moderation according to the terms and conditions can vary from platform to platform, not only due to their granularity of the definitions of the terms and conditions themselves but also to the existence of rules in the first place. As mentioned above, the category `uncategorized´ is used to include novel forms

and need of moderation or cases that might be hard to classify in the corset of the other categories.

1.2. Transparency Reporting Quality Categories

This section describes the proposed quality categories. This taxonomy of transparency reporting categories provides a selection of key reporting concepts that can be used as a comparative framework to map different transparency reporting standards in a comparative analysis. These quality categories include the concept of **Contextualization, Comparability, Vulnerability to Manipulation, Provability, Machine-Readability and Processability**. Based on previous research and literature these categories combine important elements for transparency reporting.

Contextualization

This quality category is one of the categories proposed within this analytical framework to evaluate the transparency reporting quality that claims that transparency reporting only can be meaningfully evaluated when contextualized for different purposes and recipients. Transparency reporting for online services demands also a comparatively high level of systems thinking (Douek 2022) that combines a variety of roles and user needs within a complex process. This complex process (Clune and McDaid 2023) however is the object that needs to be presented in transparency reports under the DSA. Such need for contextualization can be found e.g. in comparing notice action mechanisms of online platforms, their design and usability (Wagner et al. 2020a). For example, if content can be flagged under two systems (a contractual system to flag content according to the terms and conditions of a platform and a legal notice action mechanism as demanded by the DSA in Art 16 or the NetzDG within a Member State) the effort for reporting under such a notice action system is a relevant aspect to understand the reported numbers indicated in the numbers of the German Transparency Reports under the NetzDG (A. P. Heldt 2018). Understanding the notice action mechanism is an important concept also for user reporting (Crawford and Gillespie 2016), and can be designed in confusing ways, e.g. through the use of dark patterns (Wagner et al. 2020a) referred in the DSA to in Art 25,or provide transparency (Flyverbom 2015) and visibility (Sontheimer, Schäfer, and Mandl 2022).

Such design decisions also include considerations for meaningfully reporting on Art 20 and Art 21 DSA. The text provided to guide the user through such a reporting process is an element that must be contextualized within transparency reporting as much as the User-Interface or the notice action mechanism itself. We found that the text provided under the NetzDG was more complex in comparison to the platform's own community standards (Wagner et al. 2020a). This element is also important regarding the demands posed on platforms according to Art 14 DSA and their terms and conditions regarding the complexity of the text. This need

for understanding the rules also recurs to the concept of contextualization that only can be fulfilled if the reporting process is comprehensible to the users in detail and therefore also an important element for deterrence of the rules applied through content moderation (Pan et al. 2022) as much as fairness and trust in the service quality (Ma and Kou 2022).

Such reporting also must address specificities of individual design choices of platforms. These design choices can bring challenges to content moderation as (Zornetta and Pohland 2022) address, or be unique in their form or function (Juneja, Rama Subramanian, and Mitra 2020). Reporting in a contextualizable manner may include additional context needed to evaluate numbers or audit decisions (Tiedeke et al. 2020). It is also important to consider that unified approaches might need further contextualization regarding their community standard usage in specific regional contexts (Shahid and Vashistha 2023). This need for the contextualization of information also echoes in the domain of political content, advertising, and reporting. Contextualization not only might be needed for the content itself but also for specific moderation decisions and domains (Alizadeh et al. 2022).

Comparability

This section provides insight into why the concept of comparability is such an important element for the quality of transparency reporting. It needs to fulfil a meaningful level of human-interpretable comparability components within the transparency reporting process and should be machine-interpretable and processable to ensure comparability across different VLOPs and VLOSEs in an automated and easy manner. Comparing legal norms and compliance standards, however, also may pose an obstacle for comparability (MacCarthy 2020b) as technical details might do. This also includes the different strategies and processes of content moderation (Buckley and Schafer 2022), especially remarkable in the context of the COVID-19 pandemic and the differences in moderation approaches and tactics of online platforms to deal with such a novel moderation question (M. C. Kettemann and Sekwenz 2022).

Furthermore, languages are an important element of the quality of reporting practices for the DSA domain (Ye et al. 2023). Comparability is a key concept for providing context to different unique designs and functionalities across platforms and services (Wang et al. 2023). Comparing governance structures within platforms may also be a valuable insight into the hierarchy of content moderation that could provide comparable metrics on teams' diversity, size and location, or language background and capacity (Ahn, Baik, and Krause 2022). Comparing policies among platforms is a detail that can provide a better understanding of the way a platform regulates behaviour, to what degree, and under which conditions, including which sanctions for user behaviour a platform might foresee or leave unregulated (Einwiller and Kim 2020). Error (prediction) rates may describe specific content types, community standard content moderation, accuracy, or user behaviour and sanctions on the platform (Song et al. 2023). Besides, comparability is especially relevant for reporting illegal content to ensure transparency and reporting quality across platforms.

The class or selection of norms that might fall within a member state under “illegal content” according to Art 3 lit h DSA should follow similar norm classes and should be supported by the content moderation source indicated as proposed in this report, for all platforms that have reporting duties under the DSA. Therefore, the class “illegal content” might have a harmonizing force across transparency reporting of platforms compliance practices and their meaningful contextualisation across services. Counting violations, however, is a reporting detail that needs to be clearly set out to not make transparency reporting meaningfully comparable (Keller and Leerssen 2019).

Vulnerability to Manipulation

This quality category considers the only logical incentive to misrepresent, over or understate interpretable and complex messages and insights, the occurrence highlights, of skewed design practices, data visualization, bias or the depth of the information provided are essential elements to understanding the risk of transparency reporting (Christensen and Cornelissen 2015). Therefore, this section provides insight into measures and techniques as well as anticipative approaches to counter the challenge and the vulnerability of manipulation (Flyverbom 2016b). One approach to deal with the potential manipulation attempts might be e.g. the domain of political manipulation (Ferrara et al. 2020); for which, to a greater or lesser degree, technical solutions might overcome the risk (Flyverbom 2016a). While certain techniques such as hashing might be useful for coordination and reporting purposes (Son, Byun, and Lee 2020; Steinebach, Liu, and Yannikos 2012; Westlake, Bouchard, and Frank 2012; Yannikos et al. 2013), there also might be challenges so accurately capturing, e.g. all variants or notions of a problematic video on an online platform manifesting itself in skewed or inaccurate reporting (Hoffman 2010).

Furthermore, users themselves might game the reporting numbers (Zhao and Chen 2023) or make their behaviour harder to understand and contextualise through reporting efforts and moderation attempts (Zhao and Chen 2023). Furthermore, the DSA defines rules for the moderation of recurrent malicious user behaviour, such as continuous violations against the platform's rules or the incident of misreporting under the action takedown mechanism according to Art 23 DSA. Which should be indicated in the reporting structure including details about malicious user behaviour as well as the understanding (Kaminski 2020) that such reporting has many challenges such as user privacy (Llanos 2021a), the complexity of the visualization (Flyverbom, Madsen, and Rasche 2017) of such malicious user behaviour or the unstructured patters such details might bring to reporting and measurement and challenges. Furthermore, vulnerability to manipulation can be important for the quality of comparability efforts of the reporting results (Bradshaw and Howard 2017). Furthermore, the desire to report certain numbers that might flatter the platform might be another obstacle for the quality category at hand. This may be the case when companies claim to remove certain amounts of hate speech to a claimed degree of quality of speed (Giansiracusa 2021). This

quality category also is important for the purpose of reporting under the DSA according to Art 34 and Art 37 DSA and their addressed internal and external risk assessments.

Provability

The quality category 'Provability' stresses the point of the need to provide proof of claims that demonstrate reasoning and convincingly can deliver evidence (Michener 2019). Such demand for evidence in transparency reports, however, must target two forms of reporting – a form of reporting providing facts to humans and forms that provide evidence in an automatically comparable as well as machine readable and interpretable way.

Such measures and actions to provide proof and underline claims made can be supported through various technological solutions and could for example be implemented with technologies like the blockchain (Niu, Gao, and Zhang 2023). Such details that have to be proven however might include various privacy sensitive detail like information about marginalized users within an online community (Thach et al. 2022), the reporting on generative AI content and watermarking (Kalker, Haitsma, and Oostveen 2001), or information on human content moderators (Katsaros, Kim, and Tyler 2023). Providing proof about moderation decisions like deplatforming or impermanent suspensions as described in Art 17 and 23 DSA is another question empirically based reporting has to address (Myers West 2018).

Furthermore, evidence that might require regular updates on a current content moderation status (like suspensions or deplatforming) that might stand in contrast to the overall transparency reporting time frames can be seen as another important detail of provable transparency, and therefore, should also be referred to in the reporting times in an illustrative, understandable and user-centred way. These demands bear the prerequisite of being able to provide such proof through the systems in place (Yannikos et al. 2013) and might call also for potential costly adaptation of the current reporting system for human resources and content moderation processes and decisions alike. Additionally, the need and impulse set by regulation to create such evidence in the first place is a continuous process that has to be evaluated according to the circumstances at hand. By following concepts that include unique IDs for pieces of content or flagged accounts or posts, regulators and users can better understand the process of content moderation (e.g. indicating at what point of the process described in the internal complaint handling system a piece of content is at the moment) and the concept of process-traceability and provable accountability is supported (Hovyadinov 2019). Additionally, the heuristics of proof used to evaluate such claims are an additional detail of the concept of provability of reporting claims and can include different dimensions of argument and aspect (Tiedeke et al. 2020).

Machine-Readability and Processability

As mentioned in the quality category before, reporting has two main targets for understanding and testing the reporting claims – humans and machines. In order to automate tests, audits, and visualisations for the purpose of transparency reporting information has to be machine-readable (Lakens and DeBruine 2021). By being able to automate parts of the analysis of reported data claims made in the transparency reporting process can be proven and understood (Flyverbom, Deibert, and Matten 2019). Furthermore, big data reporting needs are demanding the technical capacity to audit such large amounts of data and the human expertise to make sense of these data based reports (Flyverbom and Murray 2018).

2. COMPARISON OF EXISTING TRANSPARENCY FRAMEWORK CATEGORIES BASED ON REPORTING CATEGORIES

This section describes how the existing transparency framework norms and categories can be compared to the developed quality categories. This comparison focuses on the transparency-norm classes and fleshes out the quality categories and transparency-norm classes connections, similarities, and gaps. By better understanding how transparency reporting processes are performed across norms this section provides insight into good practices and structures of processes and measures mapped through the quality category and the transparency-norm classes. Furthermore, this section provides a template to compare the different transparency-norm classes that combine the quality categories with specific reporting times for different reporting category (see Annex).

2.1 Overview of Standards transparency-norm classes

This section discusses the individual transparency-norm classes and their comparative mapping structure across quality categories per reporting category, indicating the reporting times per quality category and describing their specific details in the proposed order, referring to their naming (e.g., 1.1. DSA – obligatory legal transparency-norm classes) to the five classes (obligatory legal transparency-norm classes (1.), voluntary reporting transparency-norm classes (2.), and Code of Conduct of transparency-norm classes.

The transparency-norm classes are presented in the following order:

- 1.1: The Digital Services Act (*Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and Amending Directive 2000/31/EC (Digital Services Act) (Text with EEA Relevance)* 2022);
- 1.2: Regulation (EU) 2021/784 on Terrorist Content Online (*Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on Addressing the Dissemination of Terrorist Content Online (Text with EEA Relevance)* 2021);
- 1.3:The German Network Enforcement Act (NetzDG) (*Gesetz Zur Verbesserung Der Rechtsdurchsetzung in Sozialen Netzwerken – Network Enforcement Act* 2017a),
- 2.1:OECD Terrorist and Violent Extremist Content (TVEC) Voluntary Transparency Framework (OECD 2021);
- 2.2:The Santa Clara Principles on Transparency and Accountability Around Content Moderation, and associated implementation toolkits for advocates, companies (‘Santa Clara Principles on Transparency and Accountability in Content Moderation’ n.d.);
- 3.1:Relevant codes of conduct with transparency measures, such as the Code of Practice on Disinformation (‘European Economic Area (EEA) - Code of Practice on Disinformation’ 2023);

After the mapping of transparency-norm classes and quality categories are indicated per category the quality level is specified by considering the benefits and drawbacks per transparency-norm class. The section `2.2 Comparison of standards based on quality category´ propose an initial structure for linking the transparency-norm classes and the quality categories in a comparative way.

2.2 Comparison of Standards based on Transparency Reporting Quality Categories

This section compares the transparency-norm classes presented above under the quality category to indicate quality levels. This section provides insight into the quality levels by combining reporting category and quality categories in a visual form (see Annex) proposing supportive structures and practices. This includes the indication per transparency-norm class and the correlating risks and vulnerabilities.

By using the suggested table, the comparison of each transparency reporting framework can be matched according to the proposed quality category. By better understanding the interplay of the different transparency reporting advantages and drawbacks, this section provides quality levels and indicates a higher-level view on the represented classes and quality levels at hand. Additionally, this section influences the recommendations and best practices proposed in the following section.

The Digital Services Act

The DSA is the first European Regulation to harmonize reporting standards across the Member States. This regulation differentiates in the depth of the reports, according to the set of rules that apply to `intermediary services´, and `hosting providers´ (see Art 15 DSA), `online platforms` (see Art 24 DSA) and VLOPS and VLOSEs (see Art 42 DSA). Where the transparency reporting obligations differ according to platform category in a proportionate according to the role of society such a service offered might play. Such differences in reporting obligations can e.g. be addressing the internal complaint handling systems or information about human resources and content moderation. This means that VLOPS and VLOSEs have the most rigorous detail to provide.

The reporting duty the DSA foresees is annual in its basic reporting form (Art 15 para 1 DSA). This reporting duty comprises the numbers about the received orders from Member States which shall be reported in line with Art 9 and 10 DSA. This information should, furthermore, be broken down to the notices received and classified by the (alleged) type of illegal content. Including the median time span needed between receiving the order and confirming the receiving of the issued order to the requesting authority – taking a moderation action in line with the order (see Art 15 para 1 lit a DSA).

Furthermore, the DSA demands from hosting services information about the number of notices that were submitted through the notice action mechanism according to Art 16 DSA. This information should be categorised according to the allegedly illegal content category, the number of notices received that were issued by Trusted Flaggers and the moderation action taken on these requests. Specifying the notices to legal and terms and conditions violations, the number of notices that were supported by automated means and the median time between the receiving of the notice and the moderation action taken (See Art 15 para 1 lit b DSA).

Additionally, intermediary services shall provide information about the overall content moderation process. This includes information on the use of automated means (e.g., Artificial Intelligence (AI) tool support), the education and training provided to human moderators, the numbers on the content moderation action providing insight into content availability, visibility, accessibility, and other restrictions of the service at hand. Specifying the type of violation (law or terms and conditions), the method of detection (e.g., human, or automated) and the content moderation action applied (see Art 15 para 1 lit c DSA).

Besides, intermediary services shall provide insight into the notices received through the internal complaint handling system in line with Art 20 DSA (See Art 15 para 1 lit d DSA). This includes reporting on the decisions taken in the complaint-handling process, the median time such a decision would take and the numbers on reversed decisions (decisions that changed the content moderation status due to a notice brought through Art 20 DSA). Finally, Art 15 demands the provision of information about the automated means used in the content moderation process indicating the concrete purposes of their use, information on their accuracy, and their "possible error rate" and measures to safeguard and protect the automated content moderation process (see Art 15 para 1 lit c DSA).

It should be mentioned here that the information on the automated means in the content moderation process is complex in function and use. Therefore, we propose information about the target content an automated tool is built for (e.g., video content of violent protests) and their accuracy rate on average and target content. This allows us to gain a better understanding of the complexity and interplay of tools and features used in the content moderation process. Furthermore, we support a more granular way of information about the "possible error rates" with respect to false positives and false negatives and propose the inclusion of information in line with the systemic risk assessments (see Art 34 DSA) that indicates a possible risk of the false positives and false negatives specifically addressing each risk mentioned in lit a-d as an individual indication for transparency reports in question.

Additionally, Art 24 DSA provides rules for online platforms and their transparency reports. Online platforms shall indicate information about the out-of-court dispute systems (see Art 21 DSA) on the number of notices received about the disputes, the decisions taken on the cases in question, the median time between the notice and the settling of the dispute as well as the number of cases in which the online platform agreed with the decision body and

implemented the proposed moderation action (see Art 24 para 1 lit a DSA). Furthermore, the provision demands information on suspensions according to Art 23 DSA. This might e.g., include cases of deplatforming of user accounts and their legal implementation and boundaries per Member State (M. Kettemann, Rachinger, and Sekwenz 2022). The reported information on the suspensions should contain details on the "suspensions enacted for the provision of manifestly illegal content, the submission of manifestly unfounded notices and the submission of manifestly unfounded complaints" according to Art 24 para 1 lit b DSA. Art 24 also specifies biannual reporting duties in paragraph 2 for online platforms on their average monthly user numbers to determine their status as VLOPS or VLOSE. The online platforms in scope also should follow the reporting duties in line with Art 17 and the demanded SOR.

For VLOPs and VLOSEs additionally, the reporting obligation according to Art 42 applies. According to this paragraph, the reporting time frames for VLOPs and VLOSEs are biannual, and the reports should be published in at least one language of the Member States (see Art 42 para 1 DSA). Furthermore, the reporting details should also provide information on the human resources used for content moderation. This includes details about the human resources per Member State's language, reporting mechanism according to Art 16 DSA, the internal complaint handling system according to Art 20 DSA, and information about the Trusted Flaggers according to Art 22 DSA (See Art 42 para 2 lit a DSA). Furthermore, information should provide insight into the linguistic expertise of the content moderators, as well as educational measures like training and information on the services provided to support the content moderators in their job (see Art 42 para 2 lit b DSA). Besides, the transparency reports of VLOPs should provide information per Member State language about the automated systems used within the content moderation process.

VLOPs and VLOSEs furthermore have more detailed reporting obligations with respect to Art 24 para 2 DSA and demand insight into the average monthly recipients of the service – the user numbers (see Art 42 para 4 DSA). Finally, VLOPs and VLOSEs must submit to the Digital Service Coordinators (DSC) the audit reports of Art 37 DSA which are external audits testing the content moderation systems also in line with the internal audits of Art 34 DSA to make the audit reports publicly available within a time span of three months. These audit reports shall provide details on the internal risk assessment of Art 34 DSA, the mitigation endeavours undertaken by the VLOPs and VLOSEs according to Art 35 DSA and information about the consultations needed for the creation of the risk assessment and mitigation measures. The submitted information should include the external audit report as well as the audit implementation reports.

If publicly available information is containing confidential information Art 42 para 5 DSA includes a passage that allows for the exclusion of certain information. The DSA however demands significant vulnerabilities to be affected by the disclosure of information and includes areas of harm that would have to be influenced, as well as the issuance of a statement of reasons to not include the information in the transparency reports.

Contextualization

The DSA tries to answer the reporting through various ways, taken together the different reporting sizes and scopes. This logic also follows the reporting duties at hand that grow more granular in relation to their reporting category. As indicated above the DSA collects various data and information about the content moderation process at large. This includes the internal and the external content moderation process. Where the internal process could be the reporting under Art 42 DSA about the human resources of the VLOP or VLOSE, whereas the external process might be transparency reporting about the notices received according to Art 9 and 10 DSA agreeing to Art 15 DSA. By taking together external and internal structures the DSA attempts to cover the content moderation process at large. While numbers about the moderation process alone cannot attain the granularity needed for meaningful transparency reporting the first question about the reports is the shared understanding of terms and definitions needed for the reporting process, as well as the counting of variables within this reporting domain.

In this line, Daphne Keller points out the questionable definition of "an item" in transparency reporting (Keller 2023). Keller gives the example of a post that includes an image and poses the question – should this violation with embedded images be counted as one violation or two violations? This lack of context can make the conclusion drawn from the data harder or even impossible in a cross-platform comparison if not coherently followed.

We, therefore, propose the inclusion of such information in a more interactive format indicating the location of the violation in the mock-up of the UI. This can help to contextualise the numbers and better understand the problem areas of content moderation per platform. Comparing the reported numbers in this way of visualisation, however, could be challenging. Furthermore, reporting on the environment of the content in question would be another important source of information. This means the indication of information in the transparency reports that describe the effect of the aligned content on the platform. This could include the creation of a new community standard passage, clause, or explanation as well as a description of the neighbour content (likes of the content, comments etc.).

Comparability

As mentioned before, counting violations for transparency reporting is a key question for coherent and meaningful transparency reporting. Daphne Keller points out the difficulty of counting "the number of notices received, or the number of items reported?" (Keller 2023). We argue that both numbers could help to contextualise transparency reports across platforms and make their reporting better comparable if transparency reports indicate the received flags and the content items flagged in question. This could also help to understand the load of flags received in a platform comparison (e.g., YouTube vs. TikTok) compared to the amounts of content (e.g., videos flagged) perceived as a violation of some rules in the digital space. Widening the reporting frame for creating better comparability across platforms can

create shared reporting structures that interplay with the audits in Art 34 DSA and 37 DSA, and influence the creation of the mitigation measures according to Art 35 DSA and the SOR according to Art 17 DSA.

Vulnerability to Manipulation

Even if misreporting does not have to be intentional, the section above addressing the challenge of counting, and the shared understanding of definitions or specificities across platforms may pose a risk to the accuracy of transparency reports for the DSA. This perspective is shared with Keller who stresses the point of multiple violations per content (Keller 2023). This counting and reporting of multiple violations per content can be easily illustrated with the example of a video that uses music that violates copyright law, and text included in the video that would fall under the terms and conditions and the community standards of the platform as a violation of a disinformation category. While the first is a legal violation the second is a mere contractual one. If only one of those violations is logged in the system by the automated moderation system, or the human moderators (internal or external) can skew the reporting numbers, and e.g., underreport the systemic risk of disinformation.

We would like to echo this concern and have used in our previous research, several violation reasons possible per content under examination. Including terms and conditions reasons, as well as legal violations (Wagner et al. 2021; Kübler et al. 2023; Tiedeke et al. 2020). In our previous research we therefore also provided coding template structures that reflect the hierarchy of law by naming reasons for international violations, criminal law violations, media law violations and contractual violations (terms and conditions reasons). We would like to speculate on this point even further and stress the internal hierarchy of content moderation decisions in relation to work contexts.

How should a violation be counted if the automated moderation system has flagged a piece of content (false negative) and an external moderator is reviewing the AI's decision and decides that the content in question is not violating the platform rules or applicable law and if (e.g., due to quality control reasons) internal moderators take a second look at the piece of content and find that the content in question is not violating the reason provided by the automated system but it is violating another policy and because of this reasoning moderates the content in question? How should this use case be measured and reported? We, therefore, propose to also indicate the content moderation hierarchy in the transparency reports that show a timeline of moderation to make sense of the moderation decisions taken as well as understanding the actors of the process and their underlying power structures. We also support the inclusion of information that helps to analyse content with multiple violation categories to better understand the intertwined narratives, accounts and speech affected by moderation.

Provability

The modelled and presented domain of content moderation a platform provides with their transparency reports however needs to provide evidence of the status quo. As mentioned above contextualising the reported numbers with additional information is an essential need for meaningful reporting. Therefore, the strengthening of contextualisation and comparability can also help to increase provability of the reports overall. However only with the indication of clear guidelines the challenges the NetzDG transparency reports indicated can be mitigated. We therefore advise timely reporting and in cases that include the measurement of median time spans, we furthermore propose to include more granular information and reporting times that better capture the reporting process.

One example of such extended reporting times might be the reporting about suspensions or deplatforming or demonetisation of user accounts. Transparency reports should provide provable evidence of how the accounts might be suspended. This could mean provide reporting of the features limited or blocked broken down to moderation actions. Such moderation actions should indicate the reporting times and the part of the service restricted. The information and context of such a restriction might provide insight into entire blocking of the service (the recipient of the service is not able to log in), limited access to the service (the user can log into the service but cannot post publicly but still can receive private messages, make changes in the profile etc.), or increased monitoring of the account (community standards violations automatically are reviewed by a human moderator). Providing such evidence of suspension time and effect should increase trust in the moderation system in place and support understanding of the rules on the platform, as well as to create an insight into moderation actions with longer time spans.

Machine-readability and processability

According to the DSA transparency reporting should be “publicly available, in a machine-readable format and in an easily accessible manner” offered to the public according to Art 15 para1 DSA. Also, Art 24 para 5 DSA refers to the indication of the decisions according to Art 17 SOR in a machine-readable. This information, however, should take the balance between reporting and the privacy of user data into account and should not contain personal data. Providing coherent reporting standards for good machine-readability and processability is a key feature for the quality category provability as well as for comparability across platforms and other report submitters.

Regulation (EU) 2021/784 on Terrorist Content Online

The Regulation on Terrorist Content Online (RTCO) aims to harmonize the rules on terrorist online content and its dissemination online (*Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on Addressing the Dissemination of Terrorist Content Online (Text with EEA Relevance)* 2021). It demands from hosting service providers

`reasonable and proportionate´ duty of care regarding the dissemination of terrorist content and deletion or blocking of such content (see Art 1 para 1 lit a RTCO), and addresses Member States to issue protection for human rights and highlights freedom of expression among them and to act within these legal principles to build suitable safeguards to:

> "(i) identify and ensure the expeditious removal of terrorist content by hosting service providers; and (ii) facilitate cooperation among the competent authorities of Member States, hosting service providers and, where appropriate, Europol."(*Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on Addressing the Dissemination of Terrorist Content Online (Text with EEA Relevance)* 2021, 172:11).

According to the Regulation `terrorist content´ is understood as content that glorifies terrorist acts, the commitment or contributions of offences, the participation in activities of a terrorist group, and providing of instruction on the creation of explosives or firearms (this might also cover 3-D printing manuals or models of firearms), and threatening to commit such offences according to Art 2 para7 RTCO. It should be noted that the RTCO has received intense criticism by numerous civil society organisations for undermining fundamental rights protections online (Pírková 2021).

The Regulation also explains who can issue removal orders according to Art 3 RTCO. According to this paragraph, the competent Member State authorities can issue removal orders to hosting providers. The scope of the removal order should cover terrorist content according to Art 2 para 7 RTCO. A removal order should invoke the moderation action of the hosting service provider and lead to the deletion or geo-blocking of the content within all Member States (see Art 3 para 1 RTCO). The RTCO also includes an exemption for cases of emergency for the issuance of removal orders according to Art 3 para 2 last sentence.

According to Art 3 para 4 RTCO the hosting service provider has a time frame of one hour to act in line with the received removal order. Such removal orders should contain the following minimal reporting requirements: information on the competent authority issuing the order (see Art 3 para 4 lit a RTCO), a statement of reason about the consideration underlying the suspicion or decision and a reason according to Art 2 para 7 RTCO (see Art 3 para 4 lit b RTCO), a URL and further needed information on the content in question to find and remove it (see Art 3 para 4 lit c RTCO), a reference to the legal source that is violated (see Art 3 para 4 lit d RTCO), "the date, time stamp and electronic signature of the competent authority" (see Art 3 para 4 lit e RTCO), information on the redress procedure (see Art 3 para 4 lit f RTCO), and the information to not disclose the removal order in line with Article 11 para 3 RTCO (see Art 3 para 4 lit g RTCO).

The hosting service provider in the second step has the obligation according to Art 3 para 6 to inform the issuing authority about the moderation action of the (alleged) terrorist content in question. The following paragraph of the Regulation additionally foresees a clause addressing problems with acting on the removal orders demand caused by force majeure or other de

facto reasons and stipulates an information duty for the hosting provider to the competent authority.

Art 5 foresees rules on how the hosting service provider shall cope with terrorist content and its reporting. Furthermore, Art 6 of RTCO provides information on how data should be dealt with for reasons of control and proof for six months starting with the moderation action performed. In paragraph one, the RTCO a preservation duty of the content acted as terrorist content must be preserved due to testing the decision by the judicial or administrative review process or because of "the prevention, detection, investigation and prosecution of terrorist offences" according to Art 6 para 1 lit b.

According to Art 7 of the Regulation hosting service providers have transparency reporting obligations. These cover information about their terms and conditions and other policies in place that cover terrorist content. Additionally, the Regulation demands information on the functioning of specific measures and automated tools in use that address terrorist content and its dissemination (see Art 7 para 1 RTCO).

Besides, hosting service providers must publish transparency reports before the first of March each year according to Art 7 para 2 which is available to the public. The transparency report reporting obligation includes: details on the hosting provider's measures in place for the discovery and moderation action (deletion or geo-blocking of the content) (lit a), information about how the hosting service provider deals with re-uploading and sharing of terrorist content and how automated tools are used in such processes (lit b), the numbers about how many pieces of content were deleted and geo-blocked, the numbers about content that has not been disabled due to the removal problems described above or according to Art 3 para 8 of the Regulation including the grounds such exemption is based on (lit c), the number of complaints according to Art 10 of the Regulation as well as their final decision (lit d), the number of the administrative or judicial review system (lit e), information on the numbers of reinstated content due to the administrative or judicial review process (lit f), and the numbers on reinstated content due to complaint by the content provider – the user (lit g).

Contextualization

Additionally the RTCO demands context on the use of automated tools to evaluate the effectiveness and their proportionality (*Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on Addressing the Dissemination of Terrorist Content Online (Text with EEA Relevance)* 2021, 172:6). Furthermore, the Recital 24 mentions the need for the provision of information about the "human oversight and verification" capacity of hosting service providers. These tests of the effectiveness and content moderation quality "should take into account relevant parameters, including the number of removal orders issued to the hosting service provider, the size and economic capacity of the hosting service provider and the impact of its services in disseminating terrorist content, for example on the basis of the number of users in the Union, as well as the safeguards put in place to address the misuse of

its services for the dissemination of terrorist content online.“ (*Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on Addressing the Dissemination of Terrorist Content Online (Text with EEA Relevance)* 2021, 172:6).

Additionally, the Regulation emphasizes the importance of the cooperation between the Member States and the hosting service providers to put measures to mitigate the dissemination of illegal content, like the support of media literacy programs, alternative answers to content narratives and the creation of incentivised structures to minimize harm terrorist content might bring to the society at large. The Recital 2 additionally, stresses the importance of the strengthening of “social work, deradicalisation initiatives and engagement” in relation to terrorist content and its effects (*Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on Addressing the Dissemination of Terrorist Content Online (Text with EEA Relevance)* 2021, 172:1).

Comparability

The Regulation addresses the domain of comparability also with the harmonizing effect the Regulation has on this specific content category `terrorist content´. The roll-out of similar rules, therefore, should also increase trust in the service providers, as well as the user’s trust in reducing the harm of illegal content. By demanding shared reporting time frames or reporting times (e.g., one hour after receipt of the notice) the Regulation provides a hard limit on the content moderation action teams to perform the task or provide a reason of excuse. Comparing therefore not only the time frames indicated in the Reports, but also the reasons of excuse can help to better understand technical obstacles or situations hosting service providers must address in order to fulfil the compliance needs of the Regulation.

Vulnerability to Manipulation

The Regulation demands the inclusion of an explanation about how to operate on terrorist content online within the terms and conditions of the hosting service provider that sets out the „functioning of specific measures, including, where applicable, the use of automated tools“ (*Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on Addressing the Dissemination of Terrorist Content Online (Text with EEA Relevance)* 2021, 172:16). The process of making these content moderation procedures visible and concrete should mitigate different approaches towards terrorist content online. Furthermore, the Regulation addresses the problem of reoccurrence and strategic re-uploading of content in Art 7 para 1 lit b of the Regulation and their reporting about cases that used automated tools to detect such content.

Provability

If content moderation addressing terrorist content (especially the capacity of automated means), however is deemed insufficient the competent authority has the power to demand additional coping measures that address illegal content, the competent authority however should act to oblige hosting service providers to include an *ex ante* monitoring or fact-checking obligation (*Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on Addressing the Dissemination of Terrorist Content Online (Text with EEA Relevance)* 2021, 172:6).

Machine-readability and processability

The RTCO creates points of contact within the hosting service providers' structure to smoothen notice action procedures that can cope with the short moderation time frames of one hour (*Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on Addressing the Dissemination of Terrorist Content Online (Text with EEA Relevance)* 2021, 172:9). These points of contact should be built of operational teams that can facilitate the electronic removal orders submission system and act within the design process with geo-blocking or deletion within the Member States. Including a receipt or other prove of evidence for the efficient solution of the tasks.

The German Network Enforcement Act (NetzDG)

The German NetzDG has be termed as “the first regulation in the world to directly proscribe how large user-generated content platforms moderated harmful content, establishing background standards for how firms set up their complaints handling procedures, mandating a designated contact point through which the authorities could channel specific inquiries and complaints, and setting up a level of mandatory transparency reporting for platform content moderation” (Gorwa 2021, 2). The NetzDG entered into force on 1 October 2017 (*Gesetz Zur Verbesserung Der Rechtsdurchsetzung in Sozialen Netzwerken – Network Enforcement Act* 2017b) and has been the inspiration for many similar regulatory initiatives around the globe (Schulz 2022a, 292). NetzDG states that social networks that have attained more than two million users on their platform would fall into the scope of the NetzDG. If a social network is falling within the scope of the national provision “manifestly unlawful” content should be geo-blocked within Germany or deleted following a strict time frame of 24 hours if a complaint against the NetzDG is received (Wagner et al. 2020b; Amelie Heldt 2019). If the content in question is not `obviously unlawful´ the social network has a longer time frame to act on the content flagged that amounts to seven days. The NetzDG also prolongs the moderation time frame for exceptional cases.

The NetzDG is also known as the `hate speech law´, although it does not legally define hate speech but instead includes references to national law that can be associated with digital

contexts under the umbrella term “unlawful content” (Hemmert-Halswick 2021). How the term `unlawful in the context of the NetzDG in § 1 para 3 should be understood is linked to the German Criminal Code (Strafgesetz or StGB). The NetzDG selects a list of (initially) 22 criminal law provisions linked to the Member State law. In 2021 the NetzDG was amended and added now more norms to the included provisions to its initial definition of “illegal” content (*NetzDGÄndG* 2021). Some of these shortcomings of the German provision were addressed in the NetzDGÄndG – an adaption of the law. Now also the complaint mechanisms and their user-friendliness are addressed in the NetzDG in § 3 (Schulz 2022a; Hemmert-Halswick 2021). This includes according to 2 para 2 NetzDG reporting about the automated processes of content moderation, results of counter-notification procedures and more information on the terms and conditions providing a better understanding of the amount of automated moderation and human moderation. Another amendment of the NetzDG foresees researcher access under § 5a that also includes information on illegal content or deleted/geo-blocked content.

Norms listed in the NetzDG under illegal content cover the following acts:

- § 130 StGB: Incitement to hatred
- § 166 StGB: Defamation of religions, religious and ideological associations
- § 86 StGB: Dissemination of propaganda material of unconstitutional organizations
- § 86a StGB: Using symbols of unconstitutional organizations
- § 89a StGB: Preparation of a serious violent offense endangering the state
- § 91 StGB: Encouraging the commission of a serious violent offense endangering the state
- § 100a StGB: Treasonous forgery
- § 129 StGB: Forming criminal organizations
- § 129a StGB: Forming terrorist organizations
- § 129b StGB: Criminal and terrorist organizations abroad; extended confiscation and deprivation
- § 140 StGB in connection with § 138 (1) StGB: Rewarding and approving of certain offenses listed in § 138 (1) StGB
- § 269 StGB: Forgery of data intended to provide proof
- § 131 StGB: Dissemination of depictions of violence
- § 111 StGB: Public incitement to crime
- § 126 StGB: Breach of the public peace by threatening to commit offenses
- § 140 StGB in connection with § 126 (1) StGB: Rewarding and approving of offenses listed in § 126 (1) StGB
- § 241 StGB: Threatening the commission of a felony
- § 185 StGB: Insult
- § 186 StGB: Defamation
- § 187 StGB: Intentional defamation
- § 201a StGB: Violation of intimate privacy by taking photographs
- § 184b StGB: Dissemination, procurement and possession of child pornography

- § 140 in connection with §§ 176 to 178: Rewarding and approving of certain offenses listed in §§ 176 to 178
- §189 StGB (defiling memory of dead)

Normally, a piece of content is tested within the moderation process in "a two-step approach" after the (NetzDG) notification is received (Amélie Heldt 2019). First, the content is tested against the community standards of the platform – as a contractual violation of the terms and conditions – and moderated if a violation might occur. If the content is not violating the terms and conditions, the content is as a secondary step tested against the NetzDG provisions (Zipursky 2019). This hierarchy of legal or contractual violation tests can also influence user flagging and reporting for social networks and therefore mandatory reporting (Wagner et al. 2020b). Furthermore, the law states in Section 3 para 4 NetzDG that the social networks complaint handling system shall be monitored in monthly intervals by the platforms.

The NetzDG additionally provides a right to disclosure in line with §§ 14 and 15 TMG that gives the person whose rights been violated under the NetzDG relevant criminal law provisions who has a court order has the right to receive information from the social network about the person who allegedly has violated the law (Schulz 2022a; Schulz, Kettemann, and Heldt 2019). In Section 3 para 4 the NetzDG furthermore, states that "The social network's management shall offer the persons tasked with the processing of complaints training courses and support programs delivered in the German language on a regular basis, this being no less than once every six months" and demands with it training intervals in the national language, while not deciding on dedicated resources or team size.

The German law, furthermore, defines a reporting duty under Section 3 para 2 number 5 NetzDG, stating that the flagger and the user to whom the content in question belongs to should be informed about the decision taken in the moderation process including the reasoning the decision is based on. If the platforms do not follow the rules and "fail systematically" to report on their content moderation processes and their complaints handling system sanctions can follow. Such social network behaviour can be punished with high fines according to Section 4 NetzDG. Facebook was fined 2 million Euros in 2019 (*Reuters* 2019) and in 2020 over 1000 NetzDG procedures were received by the Federal Office of Justice in Germany (MacCarthy 2019). After the fine has been issued Facebook had to re-report its numbers and indicated an almost four times higher number than initially stated (Park 2020).

German jurisdiction furthermore clarified under the NetzDG that social networks may not delete content that would be classified legal under German law, may suspend user account for 30 days, and may moderate (geo-block/delete) hate speech content and corresponding user accounts even if the content in question is not deemed illegal under the NetzDG (Echikson and Knodt 2018).

Additionally, to the prominent position the NetzDG has among similar laws, the NetzDG can act as a guide to the definition of what constitutes illegal content under the Art 3 lit (h) DSA.

Furthermore, we want to highlight another Member State law and its connection to the definition of illegal content – the Austrian Communication Platforms Act which also obliges platforms to hand in transparency reports about their content moderation process (*Bundesgesetz Über Maßnahmen Zum Schutz Der Nutzer Auf Kommunikationsplattformen (Kommunikationsplattformen-Gesetz – KoPl-G – Communication Platforms Act)* 2020). In previous research, we collected an overview of the norms that might legally build the base for what might constitute `illegal´ (Tiedeke et al. 2020).

Transparency Reporting

Furthermore, the NetzDG demands that platforms that receive more than 100 complaints about unlawful content per calendar year have transparency reporting duties (§ 2 (1)). The reports should serve as a source of information to the public and "provide numbers and facts necessary in the interest of an effective impact assessment'" (A. P. Heldt 2018, 7).

These reports should be published in a biannual reporting time frame and published in German. (§ 2 (2)) describes the reporting areas for social networks in detail and should include information that described the undertakings to mitigate illegal behaviour on the platform. Additionally, information about the action takedown mechanism must be included. Details should contain the criteria used for deciding on moderation actions like geo-blocking or deletion of the content. The NetzDGÄndG additionally demands information in the transparency reports about the automated content moderation tools and systems put in place by the social networks (*NetzDGÄndG* 2021). A general description of the training data used should furthermore be indicated, as information about testing and evaluation of such systems. The reports should include information about the researcher's access provided.

The numbers about the illegal content on social networks per reporting period (half-yearly). The numbers should provide details about the complaints submitter and indicate whether the complaint was submitted by a user or a `complaint body´. The reports additionally ask for information about the complaint's reason. The new amendment to the NetzDG now also asks for information about the design ("Darstellung") of the complaint mechanism in place. It also included a passage that demands details about the decision criteria underlying the moderation process on illegal content and the testing process including the order of violation reason (terms and conditions violation or a violation based on law).

Information about human resources and the organisational of it including information on their linguistic expertise. Additionally, information about the teams responsible for dealing with complaints should be described, as well as the training of staff and the support provided for this incredibly demanding sort of work. The reports should furthermore include information about the industry association of the social networks. The reports should include numbers on the external consultation in the content moderation process. Additionally, the numbers should be included that report on the complaints that resulted in the deletion or geo-blocking of content, describing in detail the complaint submitter (user or complaint body) and after the NetzDGÄndG also the step that has led to the decision of the content moderation process and

information if the user was informed about the process whose content was questioned. The reports should indicate the number of complaints acted within the 14h time frame, the 48h timeframe, one week or the `longer´ time frame. Describing the complaint submitter and complaint reason in detail.

The reports should include information about the measures and processes in place to inform the users and complainers about the procedure. After the NetzDGÄndG the reports should also include information about the "Gegenvorstellung" according to § 3b para 3 NetzDG, a similar concept to Art 21 DSA "out of court dispute settlement". It also demands information about moderated content and anti-discrimination laws, how the spread of illegal content influences groups of users, organised structures that foster specific content spread, and measures in place that should mitigate illegal content online, including help for targets or victims. The NetzDG also asks for a tabular indication about the aggregated numbers of complaints about illegal content, the percentage of deleted or geo-blocked content, the number of the quasi out-of-court dispute mechanisms and the percentage of changed decisions. The reports should furthermore include a retrospect about the past two reporting time frames and how the differences in the numbers can be explained. The German law also asks social networks to report on information about the terms and conditions and their compliance with national provisions.

The transparency reports should be "easily recognisable, directly accessible and permanently available" according to § 2 NetzDG. The initial NetzDG transparency reports were critiqued due to their "low informative value" (A. P. Heldt 2018, 7)., the impossibility of drawing conclusions from the reported numbers (Hemmert-Halswick 2021, 421) and their misleading reporting structure (Wagner et al. 2020b). These problems within the reporting structure (NetzDG complaint or terms and conditions violation) also have financial and organizational reasons for social networks (Keller 2019; Wagner et al. 2020c; n.d.). While the NetzDG reporting is according to scope and wording close to the national provisions, the terms and conditions or community standards of a platform usually follow a global wording that should cover several jurisdictions (M. C. Kettemann and Tiedeke 2020). These similar but still different grammatical differences and legal waying however might not always be clear to the user and can burden the user if the misleading design in the notice takedown mechanism is used. Tworek and Leerssen therefore also refer to the NetzDG as the "community guidelines enforcement law" (Brignull 2019; Tworek and Leerssen 2019).

Contextualization

Publishing transparency reports alone is not enough and stressed e.g., by Heldt; who criticizes the reporting on Facebook's transparency reports and the "number of complaints [that] cannot be considered conclusive"(A. P. Heldt 2018). Only when transparency reporting is taking the context into consideration meaningful insight can be gained. What 'context' however might be in transparency reporting can take different areas of focus.

Towrek and Leerssen e.g. within this context compare the frequency of seeking “outside counsel” in the NetzDG, membership of the self-regulatory advisory bodies, processes for complain submitters and users whose content was flagged, or completeness of requests for social networks in the scope of the NetzDG into account and highlight within the transparency reports issued by Google information about the interaction needed to inform the flagger and the user connected to the content in question, which might especially be relevant to defamation cases and demanded under the NetzDGÄndG (Tworek and Leerssen 2019). According to Park “YouTube has consulted with the external legal counsel 40 times in the first half of 2018, 145 in the second half of 2018, 28 in the first half of 2019, and 2 times in the second half of 2019” (Park 2020, 36).

Heldt mentions that Facebook in 2018 had a special team dedicated to NetzDG compliance of around 65 employees and YouTube announced to employ a team of about 100 moderators dedicated to the NetzDG (A. P. Heldt 2018, 9–10). This also is made explicit in the “two-step” content moderation process that would start off with the “Community Operations” team which would first check the content under the terms and conditions of the social network and test in a second step by the “Legal Takedown Operation” team against the German NetzDG describing the differences between the German law and the terms and conditions or community standards of a social network while “It is not transparent how the teams internally assign the tasks and manage within the time based on which criteria” (Park 2020, 33). Including information on the education and training of staff (or external contractors) is not only crucial because of the complexity of moderation itself, but it can also help to understand how many resources are e.g., allocated to specific community standard classes or legal norms (e.g., are there specific training for moderators on symbols of unconstitutional organizations). As highlighted by Park YouTube indicated information about the “overall review process with the approximate timeline, the task assignment between the teams, and the practical criteria for how they assess the illegality of the content (Park 2020, 33).

If no information and context about such a case is provided a meaningful decision about the content in question might be not possible. This problem can also be captured with the mere existence of context (Kirsten/Riedl Gollatz 2018). Sometimes important content context might be already deleted and not accessible anymore. The indication of such information can however be crucial for the quality of the moderation decision. In previous research, our coding team took the existence of context into account while annotating content samples (Wagner et al. 2021). First, the annotation procedure marked whether the context about the prior conversation was available (e.g., previous posts that might determine the topic or tone of the conversation), and second information on the thread the post was attached to (e.g., naming of the thread or lead question). Including this information in the decision-making process can help to capture the shortcomings in the data and possible evaluation of the content in question.

Not only information on the organizational side of content moderation is important, but also information about the technical specificities of human content moderation stressed by the

Special Rapporteur David Kaye who proposes to include information in the transparency reports that indicate the interpretation of norms or community standard reasons and their constant development and how they are implemented including a case-law-like system that helps to better understand the actual moderation process (Kaye 2018).

Comparability

To provide meaningful insight into social network content moderation processes and NetzDG compliance across platforms reporting processes (notice action mechanisms) and hierarchies have to be standardized to a certain degree in order to compare the reported numbers across services. This problem of the diverging reporting numbers for social networks under the German law was pointed out from the beginning of the reporting duties onwards (Wagner 2020; Tworek and Leerssen, n.d.; Amelie Heldt 2019). Additionally, the differences in interpreting the NetzDG provisions have led to diverging implementation of what the law means by "supplying users with an easily recognizable, directly accessible and permanently available procedure for submitting complaints about unlawful content" under Art 2. Providing comparability across social networks on the other hand can become challenging if different services or content types are included.

These situations might lead to an "apple and oranges comparison" and might not be useful to determine NetzDG compliance after all. Prevalence of violations against a community standard or the NetzDG therefore, have to be considered with care (MacCarthy 2020b). Tworek and Leerssen provide examples from Facebook and Twitter whose complaints concentrate on the number of received complaints, compared to the reporting structure YouTube follows, which provides information on the number of content items in question (Tworek and Leerssen 2019). This stresses the challenge of standardized measurement and reporting key performance indicators. Only if the reported substance (content/counting/reporting time) and classification (content moderation source and SOR categories) are similar enough across social networks, comparative conclusions might be drawn from the reported data. Besides, meaningful reporting time or "turnaround time" frames can be an obstacle to NetzDG transparency reports (Park 2020, 33).

Prior research has shown that the written reporting of the community standards and the NetzDG can vary due to readability. We used a SMOG analysis to determine how readable the information provided within the reporting mechanisms across Twitter and Facebook (Wagner et al. 2020c, 4). Furthermore, the comparison of reporting mechanisms design could indicate quantitative measures like counting the steps needed to successfully submit a complaint across platforms, comparing the number of answer options, and visualisation, complex language or a multitude of references to attached legal articles. Our research showed that also answer options might vary in the designs of social networks that implemented NetzDG reporting structures. Including the need for user action like inserting URLs into a separate reporting window, is now also taken up by the revised NetzDGÄndG.

Another indication for problems with NetzDG compliance and reporting accuracy might be indicated if the reported numbers do not scale to the number of users on the platform (A. P. Heldt 2018, 11). Comparing moderation rates alone across platforms, however, might be as misleading as unsuitable comparisons and bear the risk of rewarding overboarding moderation action under the flag of "more moderation indicates better moderation". Additionally, the volume of complaints received might pose a challenge to comparability. While one social network might have a rather active flagging community or provide content in a language that specifically is given attention by certain NGOs or other flagging users might also influence the comparability of numbers across transparency reports. The demanded general amount of illegal content actioned on in percent on the other hand might provide some insights, e.g., how much content of the content uploaded is deemed illegal, and later classified as illegal compared to the content uploaded online overall.

Vulnerability to Manipulation

As mentioned above the reporting structure could significantly influence the reported numbers of the NetzDG transparency reports. Furthermore, the design of the complaint handling system might also influence reporting numbers regarding the location of complaint submission measures on the platform. Which also is addressed in the latest version of the law. Hemmert-Halswick promotes such features as close to the content in question and as easy to report as possible for users (Hemmert-Halswick 2021, 420). Comparing transparency reports under the NetzDG with respect to human resources and educational measures or linguistic proficiency across platforms however over the past years varied significantly regarding the depth of reporting or detail of reporting. Park highlights those differences in language expertise, educational background, team, training, well-being, fixed personnel numbers, team structure, responsibilities, or training per team unit (Park 2020, 34). These design choices can lead to a chilling effect on users to make use of the reporting tools in the first place (Wagner et al. 2020c, 6).

Provability

Provability is no final state but rather a continuous process that must be fulfilled by the reporting social networks. To fulfil this requirement of transparency timely data is a key concept of proof (Kirsten Gollatz, Riedl, and Pohlmann 2018). Data that is not existent before the reporting duty is another obstacle for the initial reporting periods for comparability and provability (Park 2020, 37). Provability might also include the detailed description of the content moderation process in an organisational understanding and how the German legal provisions are implemented in the prioritising of testing and moderation is supported by the social network ((Wagner et al. 2020c, 4–6).

Whether the classification of “manifestly unlawful” content can be meaningfully determined within these short time frames was questioned due to the complexity of testing the boundaries of freedom of speech protection for each piece of content (Wagner et al. 2021; Schulz 2022b; Tiedeke et al. 2020). Furthermore, social networks might include information in their transparency reports that proves that they have acted against incomplete complaints according to (Park 2020, 35). Such measures could ensure better protection under freedom of expression if the social networks indicate information that empirically states that they contacted submitters of incomplete complaints to provide further (needed) information on NetzDG cases.

Machine-readability and processability

Reports should be published on the social networks' own website according to section 2 NetzDG these should be permanently available. The law however does not specifically ask for concrete data formats or dashboards to be implemented by social networks.

OECD Terrorist and Violent Extremist Content (TVEC) Voluntary Transparency Framework

The OECD provides a framework for online content-sharing services that specifically target the problem domain of terrorist content. The OCED recognises the digitalisation of terrorist groups or other forms of violent speech online and mentions the “radicalisation, recruitment, dissemination of propaganda, communication and mobilisation” as key problematic behaviour (OECD 2021, 4). It therefore defines such undesired content as “Terrorist and violent extremist content (TVEC)”. TVEC proposes a reporting template within its transparency framework as well as metrics for online content-sharing services. The late version of the report highlights the legal approaches in line with the French Loi Aviva (*LOI N° 2020-766 Du 24 Juin 2020 Visant à Lutter Contre Les Contenus Haineux Sur Internet (1)* 2020), the NetzDG, and the DSA prominently (OECD 2022, 40)**.**

Contextualization

TVEC also mentions that a crisis might also increase the amount of automated handling of content moderation as in the COVID-19 pandemic and the different handling of lockdown restrictions across the globe and across platforms showed (Llanos 2021b, 5; M. C. Kettemann and Sekwenz 2022).

The TVEC report draws attention to changes in policies, wording, detection and deletion techniques, sanctions, consequences for the recipients of the service, voluntary transparency reporting the intervals of such, the content at hand or their methodologies (OECD 2021, 9). Support reporting in line with the systemic risks assessments as addressed in the DSA in Art 34 could be another interesting metric to contextualise the reported transparency numbers

(OECD 2021, 12). Twitter's indication of Trends observed in relation to TVEC content is another detail worth highlighting. This report, furthermore, stresses the information provided by Facebook and Instagram who include information on the calculation of their metrics (OECD 2021, 13). Indicating more specific information about how to calculate, what and how to measure is important information, also in the context of provability and the vulnerability to manipulation.

Comparability

Because of the recurring nature of TVEC reports since 2020 ('Current Approaches to Terrorist and Violent Extremist Content among the Global Top 50 Online Content-Sharing Services' 2020), TVEC reports are sometimes seen as a benchmark reporting over several periods (OECD 2021, 4). Keeping the coherent and recurring nature of transparency reporting in mind within the reporting architecture and design is important to also increase comparability across such transparency criteria across platforms, while keeping the balance to flexible and wide enough definitions that continue to be usable through the passage of time and development of technology. The report also mentions the risk of unharmonized rules on terrorist content handling across jurisdictions (OECD 2021, 5). Furthermore, the differences between the understanding of what is within the scope of a platform's definition of a terrorist group or extremist group might vary significantly and influence therefore, the reporting numbers (OECD 2021, 10–11). Furthermore, the reports also mention the different amounts or multitudes of violations that must be taken into consideration while comparing numbers across services or platforms – while 4chan can indicate high numbers within its reporting it could be significantly lower on platforms like LinkedIn but still may indicate a similar risk across both platforms.

Vulnerability to Manipulation

The reports highlight the importance of taking live content-sharing events into account and point to terrorist attacks in Christchurch or Halle (OECD 2021, 6). Such event-based reporting and a clear definition of a risk event might furthermore help to enhance cross-platform content moderation and design shared prevention mechanisms and emergency response techniques. Besides, the report criticizes the differences in metrics and calculations used for transparency reporting and the connected lack of comparability across platforms of such (OECD 2021, 14).

Provability

The OECD also mentions that for live streaming platforms like Twitch problems with streams might be likely for content that has been removed due to content moderation. The question is however how to count those cases within the transparency reports accurately (OECD 2021, 13). Twitch therefore uses metrics like 'enforcements' as measures counted for their reporting to still capture an accurate picture of the moderation enforcement. Another point of critiqued

is the still unknown volume of false positives and false negatives and their influence on the transparency reporting (OECD 2021, 15).

While the false positive case for terrorist content might be visualized as the restriction of legal speech, e.g., the deletion or geo-blocking of journalistic content reporting about an attack; a false negative, on the other hand, could be pictured as a case in which a piece of terrorist content (e.g., the video showing a beheading) is wrongfully uploaded on the platform and should have been moderated due to legal reasons and usually also because of a reason in the community standards. The first case (false positive) describes the case of overblocking or over deletion (Fiala and Husovec 2022). This overblocking approach therefore stands in contrast to human rights and the protection of freedom of speech (Sander 2019).

Machine-readability and processability

TVEC especially highlight the endeavours of the use of hashing techniques and digital fingerprints to detect and moderate in a platform environment violent and terrorist content (OECD 2021, 6). The report also relates to the initiative undertaken by GIFCT ('GIFCT' n.d.; Mancuso (CELA) 2020) and their Hash-sharing database and highlights the taxonomy created to hash TVEC content created by a hashing consortium (OECD 2021, 17).

The Santa Clara Principles

The Santa Clara Principles (SCP) go back to the year 2018 and their initiation by several human rights groups and experts in the field. The group designed principles that should guide content moderation and its reporting in a meaningful and accountable way (Solomun, Polataiko, and Hayes 2020, 6). According to Urman and Makhortykh the SCP can be regarded as basic suggestions on what information companies should include in their transparency reports, especially in the context of content moderation" (Urman and Makhortykh 2023, 9)**.**

These principles are Human Rights and Due Process, Understandable Rules and Policies, and Cultural Competence, State Involvement in Content Moderation, State Involvement in Content Moderation. Each principle furthermore includes guidelines on how to implement it. Furthermore, the SCP define operational principles, principles for governments and other state actors. Several VLOPs and VLOSEs have endorsed the SCP since its launch.

The SCP propose within their operational principles the reporting of the total amount of content moderated and accounts suspended, the number of appeals against content moderation decisions of the platform, the amount, or the percentage of reinstated content due to appeals, including their unsuccessful share, the amount or percentage of appeals that were "initially flagged by automated detection", the amount of ex-ante reinstated pieces of contents or accounts, information about hate speech policies detailed to targeted groups and

attributes considering the privacy rights for such user groups, as wells as reporting on crises situations and their content moderation influence such as a pandemic.

Furthermore, the reporting involving state actors should include detailed information that provides insight into each country separately. According to this scenario, reporting should furthermore include the amount of state actor requests to the platforms for content or accounts to be moderated. Additionally, the SCP advises including the identity of the state actor and the request demanded by it. Besides, the information about the flagger should set out information about the request and should specify if the notice was brought due to a court order or judicial request or by some other kind of governmental actor. Also, the differentiation of the type of violation in relation to the flag received and the action is taken (terms and conditions violation or legal source violation) should be indicated. The SCP also highlight within this case the potential indication of both reasons – which we also support.

The SCP also highlight that meaningful reporting should include information about the overall number of notices received over some period of time (TRT), the overall number of notices "traced by bots", and the overall amount of content and accounts moderated per alleged violation and flagger, like governmental actors, Trusted Flaggers, normal users or automated tools used within the process.

The SCP additionally highlight the importance of automatic tools and processes used for content moderation or within its process. According to the SCP, transparency reporting should describe `when´ within the process and `how´ within content moderation (completely or semi-automated) should be included in the reporting. Furthermore, the SCP describe the need for reporting on the `categories´ and `types´ of content for which automated tools are used. Additionally, the information about the main criteria that are underlying the decision-making should be reported by platforms. Besides, rates that indicate the `confidence/ accuracy/ success´ of the automated tools should be reported, as well as the changes (like new versions or updates) and the information about the rates for different languages and content categories.

Furthermore, the human oversight and redress mechanisms to humans should be included in the information provided by the platforms. Also, numbers about the appeal process should be indicated within the reports. This included information about content, as well as accounts and successful and unsuccessful appeals submitted to the platform. These details should also indicate the content type and the content category it is classified under, specifying (terms and conditions violations and legal violations). Finally, the SCP propose to report on the various platform databases used to support shared techniques like hashing to coherently apply rules across platforms. The SCP additionally support a transparency reporting interval of quarterly reports that are openly licenced.

Contextualization

Within the operational principle, the SCP highlights the coherent reporting understanding and "the importance of transparency in content moderation, both to users seeking to understand decisions about their own speech and to society at large. Companies should report information that reflects the whole suite of actions the company may take against user content and accounts due to violations of company rules and policies, so that users and researchers understand and trust the systems in place" ('Santa Clara Principles on Transparency and Accountability in Content Moderation' n.d.).

The SCP, furthermore, defines standards for how platforms should set out the design of notices to users and how to make the information about content moderation actions central within their platform's architecture. Such details also include the exemption of these rules and their information about such in the terms and conditions or other policies of their platforms. These details also should be taken into consideration when it comes to reporting in a holistic way. Only if users are informed about their moderation action, appeals about such can be issued. This reporting and notifying processes and designs therefore significantly influence the numbers within the transparency reports. The SCP also sets out rules for the appeal process implemented by platforms and what such an implementation could look like. Only if context to the appeal process, its design and explication to users and their reporting is taken into consideration transparency reporting can provide meaningful insight through the indicated numbers.

The SCP also support information to be provided by governmental actors about their submission of notices to the platforms. This indication of the government can help to increase meaningful context to individual content or account cases and might also increase comparability.

Comparability

Furthermore, reporting across platforms is not only led by the information provided to governments but also to the information not provided to governmental actors according to (Urman and Makhortykh 2023, 9). They compare the indication of Amazon's transparency reports and GitHub's and their information about the governmental request. We also would support such information to be included within the transparency reports. Furthermore, they provide a comparison of the transparency reporting depth of TikTok in relation to ads that were moderated, Twitter's details on 'malicious automation' or Linkedin's information about fake accounts and content classified as spam (Urman and Makhortykh 2023, 10). They also highlight a detail reported by Facebook – content that has a large number of views, or the number of potential view numbers of violating content.

Vulnerability to Manipulation

The SCP calls for reporting about the integrity of the service and includes in their definition of integrity “efforts to protect the integrity of their services against manipulative and abusive conduct“ (‘Santa Clara Principles on Transparency and Accountability in Content Moderation’ n.d.). This includes the presence of information addressing the problem of manipulation or inauthentic forms of behaviour within their policies and setting out strategies on how to cope with this risk. The SCP, furthermore, stress the lack of knowledge of the topic, and mentions the global nature of operations. Therefore, the SCP propose to include more granular information about coordinated inauthentic behaviour.

Provability

The SCP also demand a clear procedure and structure for naming and labelling of the policies and systems that underly the transparency efforts. This includes a detailed labelling for political ads that also provides information about the sponsor of the ad and the candidate or party that is financially responsible for placing it. By providing these details, the SCP furthermore stress the need for correct labelling of ads as `political´ and the need for searchability to provide reporting quality. Also, the SCP stresses the diminished value of explanations like `targeting based on user preference´ (‘Santa Clara Principles on Transparency and Accountability in Content Moderation’ n.d.)

Machine-readability and processability

The SCP support the submission of transparency reports that are machine-readable. Uraman and Markhortykh also emphasise initiatives like Lumen and the membership of such database-centred reporting across platforms (Urman and Makhortykh 2023, 10)**.**

Relevant codes of conduct with transparency measures

Codes of conduct are referred to in the DSA under Art 45 and acknowledge that there are specific risks to content moderation which could be supported by the creation of such codes of conduct (see Art 45 para 1 DSA). These to address risks are linked to the systemic risks as mentioned in Art 34 DSA and should be supported by the inclusion of platforms, civil society and other stakeholders like public authorities NGOs or Trusted Flaggers. Such codes, therefore, should provide guidance and mitigation strategies and reporting frameworks for risks addressed (see Art 45 para 2DSA).

Furthermore, the DSA seeks to define Key Performance Indicators within the issuance of codes of conduct (see Art 45 para 3 DSA). Furthermore, the DSA specifies rules on codes of conduct for advertising in Art 46 DSA and for accessibility in Art 47. The EU has taken the initiative to

issue content moderation related Codes of Conduct like the Code of Conduct on Hate Speech (*The EU Code of Conduct on Countering Illegal Hate Speech Online* 2019) or the Code of Practice on Disinformation (*Code of Practice on Disinformation* 2018).

The aim of the creation of such codes of conduct is furthermore, to create a Brussels effect and harmonize reporting and safeguarding efforts across borders – and names the Code of Practice of Disinformation as an excellent example of such an effect (Nunziato 2023).

Code of Practice on Disinformation

The Code of Practice on Disinformation (CPD) (*Code of Practice on Disinformation* 2018), here also referred to as `the code´, is providing "measures that may curtail advertising revenue and the impetus that gives to the dissemination of certain content; and encourages transparency, system integrity, media literacy and research access" (Chase 2019, 2). Its initial signatories included Facebook and Instagram, Google, Mozilla and Twitter and were joined a year later by Microsoft and includes within its scope the European Economic Area (Chase 2019, 5).

The CPD recognizes the risk of Member State citizens being exposed to disinformation and should act as a coordinated measure against it. The code defines disinformation as "verifiably false or misleading information" which, cumulatively, (a) "Is created, presented and disseminated for economic gain or to intentionally deceive the public"; and (b) "May cause public harm", intended as "threats to democratic political and policymaking processes as well as public goods such as the protection of EU citizens' health, the environment or security" (*Code of Practice on Disinformation* 2018, 1). This definition however excludes satire or "clearly identified partisan news and commentary" for the definition of disinformation (Chase 2019, 5). The initial code however did lack a "common approach" that provided measurements and other indicators of success of the code (Chase 2019, 9). These voices were heard however and led to the creation of the Strengthened Code in 2022 ('The Strengthened Code of Practice on Disinformation' 2022). This new code included 44 commitments and 128 specific measures and includes now 34 signatories. Furthermore, the code provides qualitative reporting elements and service level indicators, which act as measurements for effectiveness and implementation success. Furthermore, the strengthened Code creates a new Transparency Center that includes a permanent Taskforce. At the same time, numerous civil society groups have heavily criticised stakeholder participation within the CPD, suggesting considerable challenges in reaching consensus around the stakeholder consultation process and the substance of the CPD (AlgorithmWatch 2022).

The CPD formulates strategies for coping with demonetization, transparency and political advertising, manipulative behaviour, the emancipation of users empowering researchers, and enfranchisement of the fact-checking community.

Contextualization

The code additionally stresses the importance of the balance of freedom of expression and the protection of lawful speech for the signatories of the Code (*Code of Practice on Disinformation* 2018, 1). The signatories according to the Code recognize a multitude of challenges of disinformation, like the prioritization of `authentic, and accurate and authoritative information´ in feeds (see vii), user transparency for political targeting and advertising (see viii), or the initiatives needed against fake accounts (see v). By signing several commitments platforms can place specific practises including the scrutiny of ad placements, political advertising and issue-based advertising, integrity of services, and empowering consumers, empowering research community. And through this also recognizing location, placement and content environment overall as critical aspects of addressing disinformation.

The user empowerment is furthermore brought up in the Code within the context of `why am I seeing this ad´ and political advertising (*Code of Practice on Disinformation* 2018, 5). Due to the code platforms started to roll out a series of measures in support of the Code like the strengthened support of civil society groups and NGOs, building up electoral security centres, fact-checking partnerships across the EU addressing content in several languages, courses for journalistic expertise on the topic providing coping strategies and verification techniques, or the promotion of verified or authorised content (Chase 2019, 22).

In the line of SLI 21.1.2. we, furthermore, propose to also include the information about the verification of a piece of content as a log entry that is included in the information about the overall piece of content and should be reported in the transparency reports. The code therein demands information about "[…] actions taken at the Member State level and their impact, via metrics, of:

- number of articles published by independent fact-checkers
- number of labels applied to content, such as on the basis of such articles
- meaningful metrics on the impact of actions taken under Measure 21.1.1 such as the impact of said measures on user interactions with, or user re-shares of, content fact-checked as false or misleading."('The Strengthened Code of Practice on Disinformation' 2022, 28).

The code, furthermore, addresses the issue of verification and authentication of political ads and political campaigns and accounts. Information about views of such ads furthermore can help to contextualize, e.g., risk for political disinformation. Additionally, information about the verification process, the political accounts denied, accepted, the time frame taken between the initiative taken by the user or political campaigner would help to contextualize the ease of application and quality of verification and testing for the mitigation of risks.

Comparability

The code highlights furthermore, the importance of a holistic view on the problem of disinformation within the online environment and stresses the importance to work not isolated within this interplay of roles (*Code of Practice on Disinformation* 2018, 4). Furthermore, the Code acknowledges the importance to include strategies and control for all Member State languages. This aspect to comparability is demanding user-friendly information on policies addressing disinformation and the tools in place to flag content under such policies in place ('The Strengthened Code of Practice on Disinformation' 2022, 31). Another aspect of comparability in contrast to the NetzDG or the TVEC is the mere contractual violation nature disinformation that is usually not illegal (Chase 2019, 1). Another aspect of comparability across platforms might be the effect that content moderation of disinformation has on the different communities. This is made explicit in a case brought up by Chase, who stresses the future effects of moderation on a user account and provides the example of a downranked Facebook News Feed and the significant reduction of future views for the Feed of about 80% (Chase 2019, 19).

Vulnerability to Manipulation

The code considers the risk of human and machine-based abuse of the service and mentions examples of such "machine-based abuse" as malicious mass-flagging of content and "disclosing information that would help would-be abusers find and exploit vulnerabilities in their defences" ('The Strengthened Code of Practice on Disinformation' 2022, 31). Furthermore, information about "Coordinated Inauthentic Behavior" (CIB) is highlighted by Chase in Google's reports (Chase 2019, 19). We in line with this emphasis would also advocate for the inclusion of information about identified keywords, behaviour patterns or shared tactics in line with the qualitative transparency reporting for VLOPs and VLOSEs.

Provability

Not only the correct reporting of content moderation numbers is relevant for the quality of the reporting process, but also the submission of meaningful information about design choices and compliance implementation measures must be proven to the overseeing body. According to scrutiny of the ad placements, the Code demands signatories to put financially reasonable efforts into place to not monetise such content or accounts or promote the content of continuously violating disinformation policies (*Code of Practice on Disinformation* 2018, 4).

Proving to have such measures in place could therefore include accuracy measures (false positives, false negatives), accounts or content types of the measures that do not work well on e.g., specific Member State languages, as design elements, UIs, mock-ups, policy design (e.g., the tracking of changes over time) or advertiser option. The code foresees the information time frame of about 30 days after the changes occurred in the policies of platforms ('The Strengthened Code of Practice on Disinformation' 2022, 42). We propose to

re-use the time frame indicated in the code for the reporting times in relation to other reporting on policy changes as a good time frame of general reporting about such changes, and add an exemption to political advertising, campaigning, TVEC or crisis related content of an adequate corresponding time frame.

Machine-readability and processability

The code supports the use of the IPv6 to increase provability through the information provided that can be traced back to single users as a processable electronic mean of identification (Chase 2019, 4). Furthermore, the initiatives are taken by the Code to increase searchable databases for labelled political ads. Additionally, the code demands information about transparency to be placed at Transparency Centers which should be "publicly available, user-friendly, and searchable" ('The Strengthened Code of Practice on Disinformation' 2022, 41).

3. RECOMMENDATIONS AND BEST PRACTICES OF TRANSPARENCY REPORTING

This chapter provides recommendations and Best Practices for transparency reporting. The aim is to provide a novel standard of transparency reporting. The recommendations are based on the quality levels demanded per norm and summarized in clustered points of argumentation reflecting on benefits and drawbacks per quality category and transparency-norm class including an analysis of the benefits and drawbacks of such decisions and practices.

Contextualization

Different contexts for different roles and users

Considering that different recipients are addresses of transparency reports that have various needs is central. The parties putting together the reports therefore should also consider the different perspectives of report addresses and curate the information in a way that is meaningful for the various roles and the context at hand. **Different information receivers** should be able to gain knowledge through these reports regardless of their professional or personal background.

Different roles are involved in socio-technical processes

Furthermore, it is important to acknowledge and document the different roles included in the process of content moderation in the first place. These include, but are not limited to:

- automated tools (in-house developed or external services used) built-in the process,
- human content moderators (internal content moderators, external content moderators, dedicated moderation teams, outside counsel, linguistic expertise, hours worked in a row, location etc.),
- external flaggers (normal users, Trusted Flaggers, public authorities, etc.),
- regulators (DSC, Commission, Member States),
- other platforms or App stores.

Better understanding the **complex interplay of automated tools and human moderation** is another important quality criterion for transparency reports in our opinion.

Unique IDs for content moderation roles

Following the **unique ID approach**, we furthermore support information on the hierarchy of content moderation. This includes information about the **content moderator's position** (internal, external), **special training** (e.g., NetzDG specialist, or specialised in the moderation

of “hate speech” content in Germany) as well as the power to **overrule a prior decision** (e.g. in cases in which first an automated system takes a decision which is later changed by an external content moderator but later changed by an internal content moderator who has the same position).

Reporting on the hierarchies of content moderation

Understanding **hierarchies of content moderation** can help to understand power structures within the teams and across platforms and should help to understand the interplay between automated tools and human moderators in detail. Additionally, **visual diagrams** could help to understand content moderation procedure within the online platform. These visual diagrams would need to be created by the platforms. Third parties, like researchers and Trusted Flaggers, or DSCs and the Commission, however, can use such a visual diagram to make sense of the structure and process of the platforms. Only when the unique process of content moderation is illustrated and can be understood by researchers a meaningful quality assessment of their moderation systems in place can be provided.

Additionally, we support the inclusion of information about the **decision-making process** to **change community standards or tools** used in the content moderation process (like content moderation software for humans or the inclusion of automated tools, or the power to change the prioritisation of tools). Providing information on **which teams are included**, or if there is **negotiation power** in the **implementation** process of new rules (Fischer and Kraus 2022). This also should include the power and hierarchy information about the change of definitions and implementation of the community guidelines.

Detailed information about the human support in content moderation

Additionally, we recommend the indication about work-related details for content moderators and external counsel supporting the content moderation process. Such information should typically include details about the content moderation team’s diversity and background, target content type or category, the size of the teams and the languages spoken, as well as information about their workload (hours per team provided) and capacity (pieces of content moderated). Furthermore, in line with the current version of the Platform Workers Directive Art 6, we support the indication about how content moderators work is evaluated and monitored (European Parliament and the Council and Council of Europe 2021)

Example decisions and explanations

We also support the inclusion of a set of randomly selected examples – insofar as these are not illegal for platforms to publish online and do not infringe the right to privacy - that **illustrate the decisions** taken in the moderation process across policies. One set of examples that goes in this direction are the examples provided by Google (‘Removals under the Network Enforcement Law – Google Transparency Report’ n.d.). These examples can help to

contextualise the reporting and provides an overview for report receivers which challenges the platform's rules have to deal with.

Reporting on relevant organisational and financial changes for VLOPS and VLOSES

We also want to stress the importance of **changes within the organisational or financial structure of online platforms** as well as the impact **Mergers and Acquisitions** can have on compliance efforts and continuity (Miller et al. 2023). These should be communicated clearly and consistently as part of transparency report, to indicate clearly their consequences for content moderation decisions.

Moderation times and moderation quality

Furthermore, information about the **moderation time spans of humans** can help to understand the complex process underlying taking decisions in line with the law or the policies in place (Wagner et al. 2021). We included time frames in our measurement from encountering the content for the first time including the time needed to look for additional context in order to take a final decision as an end point. We, furthermore, propose to include information on the time needed to take the **moderation action (where internal/external/counsel)** and the time the content was up for viewing including details about the time to flag for Trusted Flaggers submitted complaint to the platforms action to moderation.

Including concrete moderation times can also help to support and protect content moderation workers (how much time does a good decision typically take?) and ensure higher content moderation quality (below which time threshold is good content moderation essentially impossible?) to create data about quality content moderation. We used in prior research the indication of perceived certainty per decision to indicate the difficulty of the task at hand and to further contextualise the subjective content moderation process.

Granular reporting for VLOPS and VLOSEs on legal and other capacities of content moderation teams

Details on the **legal and other training as well as degree of support of the staff** is key to understanding content moderation decisions. This should explicitly include psycho-social care for staff, technical and organisational mechanisms to support content moderators and other measures to promote wellbeing among staff, as well as information about staff turnover and longevity. As high levels of training of staff will typically increase costs, platforms may feel an incentive to save money on staff training and support. It is therefore crucial that transparency reporting provides an overview of measures specified in this area, to help better understand what type of training and support enable high quality content moderation, high staff wellbeing and low staff turnover. Taking these questions into consideration is likely to be important for **comparability** reasons and to ensure **meaningful reporting**.

Role-Action Timeline for content moderation

We also support the information within a **role-action timeline** that helps to better understand the moderation decisions in a process way or diagram. This could be supported by the **unique IDs** mentioned and a **user-friendly visualization** attached to the information at hand. Such a `**content moderation process diagram per decision**´ can help to contextualize reporting information overall.

Strengthening reporting on soft moderation

Additionally, we want to stress the importance of the inclusion of information about **soft moderation** within the qualitative assessment like the attachment of warning labels. This also includes information about how soft moderation measures are explained within the terms and conditions to the user. Furthermore, we support the increased reporting effort and indication of information about the explanation types, accuracies and depth e.g., provided on explainer functions provided through warning labels (Ling, Gummadi, and Zannettou 2022; Zannettou 2021). Better understanding the explanations provided to users about why certain content is presented to them is a curtail detail of the overall content moderation process. **Including descriptive soft moderation measures** therefore in the **SOR according to Art 17** DSA could furthermore help to monitor the moderation behaviour in a holistic and meaningful way that additionally supports transparency reporting through the inclusion of **coherent and understandable categories for warning labels** or other descriptive soft moderation measures.

Mock-Up User and Moderation Interfaces for Reporting Purposes

We support the inclusion of **mock-ups of the user and moderation interfaces** in the reporting of the platforms. This can help to better understand the functionalities of the platforms or services and helps to understand how the content might be presented and consumed by the user, which could be quite different across the services provided. As mentioned above transparency reports include many roles, and so do the **UIs** provided within the process. Therefore, we support a mock-up for the following user groups: the recipient of the service, **the (normal) user flagger, the Trusted Flagger, the public authorities, and the moderators (internal/external/ counsel).** Contextualising their presented digital realities helps to contextualise the reported numbers across services.

Keller additionally, highlights that information about the **authorization and testing process for verifying public authorities and trusted flaggers**. As mentioned, above the indication of unique IDs could help within this authorization process. Furthermore, **a more granular description of categories for different Trusted Flaggers** according to Art 22 DSA, indicating their reporting duty Art 24 para 3 lit a-c DSA might be beneficial to map out competencies or flagging gaps (areas of content that would or not `naturally´ be flagged by Trusted Flaggers) within certain content areas as a data sub-category. Furthermore, such a **mock-up reporting**

structure helps to better understand the **specificities of the platforms** themselves, like Pins on Pinterest, filters on TikTok or reals on Instagram.

Furthermore, Keller points out the importance of **community-moderated platforms** and their transparency reporting according to the DSA. How should such community-related content moderation actions should be presented within the transparency reporting templates could be also addressed by the **role-process diagrams** which would provide the community moderator with a **unique ID** which is indicated in the transparency reporting diagram. This would also help to contextualize the community moderation process for research and transparency purposes.

Coherent counting and multiple indication approaches to reporting about content actioned

We also support in line with Daphne Keller the documentation of: **The number of notices received per SOR category, and (per violation) and the number of content notices, (per content) to provide a more holistic view within the transparency reports** (Keller 2023). Furthermore, we support the information about the **average prevalence of violation against each community standard and legal reason per Member State** for content online. **Making different violations visible**, as well as their `blurry boundaries´ are important to understand the complex process of content moderation (Wagner et al. 2021).

Reporting on Affected Content

Additionally, we support the documentation about the **`affected content´**. This means, which content is influenced by the moderation decision in question? **An example of such an influence would be the deletion of a question** in a thread under which several answers would be included. Information that provides insight into what happened with the content posted under the moderated question would complete the picture current transparency reports provide. Furthermore, we stress the importance of the granular reporting of category areas like the disinformation in relation to the climate crises (Institute for Strategic Dialog 2023).

Reporting per content type

We also support in line with Tworek and Leerssen the inclusion of information about the **compliance rate per content type (image, audio, video),** and add the inclusion of time frames indicating the moderation time needed across content types and their underlying complexity and tools needed (Tworek and Leerssen 2019). Furthermore, we support the inclusion of information about the **public authority's data, action or information requested** providing insight on the content in request's demand and the alleged violation reasons, as well as the internal processes used to deal with such requests (Belli, Zingales, and Curzi 2021).

Reporting on accuracy rates for automated tools per content category

We propose to indicate information about the target content and the automated tool that is used in the content moderation process for such tasks. This includes the connection with accuracy rates per content type, e.g., is the hate speech classifier equipped to detect hateful speech in audio files? Indicating these accuracy rates per content category therefore should show on an average sample of content targeted how it would perform. We also support the indication about the strength and weaknesses of such tools used and their false positive and false negative rates for the average performance per content category (content moderation source) used. According to Recital 96 DSA the DSC or the Commission may request information on the accuracy of VLOPs and VLOSEs. Additionally, Art 16 para 1 lit e mentions "indicators of the accuracy and the possible error of the automated means" used in the content moderation process. Furthermore, Art 42 para 2 lit c states that "indicators of accuracy and related information […] should be broken down by each official language of the Member States".

Reporting in line with (perceived) systemic risk categories by design

We also support the inclusion of **risk reporting categories** in the transparency reporting process. We think that by aligning reporting according to audits and transparency reports better quality for both can be achieved. Furthermore, reporting e.g. on notice action mechanisms if combined with risk categories can provide more insight into the location or occurrence of risks. These should be broken down in line with the systemic risk assessment for VLOPS and VLOSEs according to:

1. Illegal content (lit a),
    a. Indicated per Member State
2. the risk to fundamental rights (lit b),
    a. Art 1 of the Charter,
    b. Art 7 of the Charter,
    c. Art 8 of the Charter,
    d. Art 11 of the Charter,
    e. Art 21 of the Charter,
    f. Art 24 of the Charter, and
    g. Art 38 of the Charter.
3. the risk on 'civic discourse and [the] electoral process as well as the public security (lit c)
4. and the risk that 'gender-based violence' poses, to public health physical and mental well-being, or the protection of minors (lit d).

Comparability

Coherent counting and multiple indication approaches to reporting about content actioned

Additionally, we support the comparison of platform transparency reporting by (Urman and Makhortykh 2023). They compare across different platforms according to the **total number of contents actioned, number of appeals, number of successful and unsuccessful appeals, number of unsuccessful appeals, initially flagged by automated detection, number of posts or accounts reinstated proactively, numbers on enforcement of hate speech policies, number on content removal during crisis periods**.

Furthermore, in line with MacCathy we support improve overall comparison of transparency reporting (MacCarthy 2022, 10). MacCathy differentiates in content moderation programs, advertising, and operation of the service. The first category differentiates between **content rules, enforcement procedures, complaint process, misuse policy, misuse warnings, explanations, redress rights, trusted flaggers, operation of program, algorithms** – main parameters, and algorithms technical detail. These variables are compared to information types: **terms and conditions, user, public report/database, auditor, vetted researcher, and regulator**. These information types also apply to the comparison of advertising and operation of the service. Advertising furthermore includes the subcategories**: content, sponsorship, total and targeted audience, main targeting criteria, algorithms** – and technical detail. Operations of service on the other hand differentiate into six categories: **average monthly users, monitoring and compliance data, recommendation algorithms, commercially sensitive data, systemic risks, and personal data**.

He also compares similar criteria in an earlier report on disclosure recommendations by **audience information type** (MacCarthy 2020a). He indicates four information type categories: **content moderation programs, reports, algorithms, and data**. The category content moderation program is quite like the equivalent category mentioned above, where it does not hold as many differentiations on the topic as the more recent publication. Reports only mention the class content moderation, **algorithms however are detailed as technical descriptions** according to MacCathy's category that differentiate **content moderation, priorisation and recommendation** details. The data category is described as **content moderation programs, political ads, content-ordering techniques, and commercially sensitive/personal**. All these categories of information types are compared to **three roles**: the **public, vetted researchers, and regulators**; describing for each role the relationship with an information type and sub-type.

Keller also mentions the problem related to **multiple counts** of copied and moderated content and provides the example of a music file that would fall under copyright provisions and moderation action could be based on such, while platforms might also use their contractual power to delete more similar copies of this file. We support the counting system that covers the requests received as **legal violations** because the platform has already knowledge of the

claim, and the objecting party would probably – if informed – file several more requests about the copyright infringement on similar pieces detected by the platform if informed.

Additionally, in line with our own coding procedure (Kübler et al. 2023; Tiedeke et al. 2020; Wagner et al. 2021) we also **support a count of each violation possible** that covers the overlap of legal and contractual violations alike. Keller also mentions the problem of **counting governmental or public authority requests resulting in content moderation action as a category that could be counted as a terms and conditions violation or removal due to the law**. We would suggest basing such requests on **legal reasons** that provide the public authorities with the power to request.

Furthermore, within the **SOR (see Art 17 DSA) the legal reason** should be provided to ground a claim in law. Additionally, requests could also be counted as terms and conditions **reasons** if there are **more granular descriptions on the groups of requests** (e.g., requests about the removal of journalist accounts, political advertising, or specific narratives) could also support a better understanding of the interest, narratives and roles for research purposes and the information of the public.

There are two good practices that could avoid these problems of multiple counting and provide more robust transparency. The first would be to cross-reference the transparency reports with the statements of reason database items included in the respective category. To implement this, platforms could provide for each category an annex per transparency category, providing a full list of links to statement of reasons entries which match that transparency category. The second is to attach unique IDs to types of content on platforms that keep reappearing, for example terrorist content that keeps being reuploaded across the platform every time it is removed. By providing a unique ID for that content and using it systematically in transparency reports and the statement of reasons database, it would be possible to avoid some versions of multiple counting described above.

To be clear, both of these practices might be overly burdensome for smaller online platforms and are mainly considered with large online platforms in mind. At the same time, once correctly technically implemented, their implementation would likely not be unreasonably burdensome for a platform with the size and resources to setup a robust transparency workflow that is striving to implement good practice around implementing DSA transparency. Linking transparency reports and statements of reasons as well as making transparent multiple counting in transparency reports would provide a much higher degree of transparency and comparability and should thus be seriously considered when implementing DSA transparency requirements.

Comparing overblocking biases and incentives

Additionally, Fiala and Husovec demonstrate in a laboratory experiment the effect overblocking bias has on the reporting and appeal behaviour between users, platforms and

resolution bodies (Fiala and Husovec 2022). Taking these dynamics into account in the comparisons across platforms is also important in a comparison between VLOPs and hosting services, or online platforms, regarding to organization, mitigation of negative effects and financial effects of such overblocking biases.

In their publication, they stress the importance of user empowerment and their often-neglected role in the information chain about the content moderated. We support the inclusion of information on how users are empowered and involved in the content moderation policy creation broken down to Member States. These reporting details might increase the quality category of Comparability across national content topics and platforms. Furthermore, such information and proactive action could help to bridge gaps identified by recipients of the service. Such an option might include reporting about the possibility for the user to inform the online platform or VLOP/VLOSE concerning a needed keyword to be included in the Art 17 DSA SOR. Providing such a tool could therefore strengthen user empowerment, the collection of moderation policies, categories and problems from a user perspective and could be addressed in a coherent and shared manner.

Vulnerability to Manipulation

Design of the notification, complaint handling, and dispute resolution systems

Moreover, the structure and definitions referred to in Art 20 DSA that indicate details about the **internal complaint handling system** should be taken into account for the transparency reporting structure including a **Unique ID** for the reporting process. Therefore, **Art 24 para 3 lit b** DSA **(information about suspensions, such as deplatforming)** should be specified in a similar definition as the proposed list of content moderation source.

Systemic reporting on malicious attacks and moderation action

Additionally, we support the **coherent counting** and **reporting on user accounts** and malicious user behaviour in line with **Art 23 DSA.** This furthermore might include information on targeted accounts or topics in line with **malicious user attacks**. By providing such behaviour information within the transparency reports internal **risk assessments** (see Art 34 DSA) and **external audits** (see Art 37 DSA) can be supported. Furthermore, the **validity of violation claims** has to be tested and can be a vulnerability to the reporting quality (Fiala and Husovec 2022, Urban and Quilter 2006). As Fiala and Husovec also point out is the potential discrepancy of reported error rates and “procedural `mistakes´” and their relationship between financial incentives and the solution of harder (legal) problems at hand (Fiala and Husovec 2022, 2).

Agreed on accuracy rates for the use of automated tools

We recommend developing standards for common accuracy rates for VLOP or VLOSE to use an automated tool in their content moderation. These accuracy rates do not have to be static and can vary from content type to category moderated, but their indication can show

potential vulnerabilities of automated content moderation as well as areas that need stronger human support to reach a better level of content moderation quality. Furthermore, this helps to establish a common baseline of accuracy that can be used in comparisons for platform moderation tools in place and could help to detect risky uses of inaccurate or not well-functioning (support) tools in place and can be utilised in internal and external risk assessments.

Provability

Provision of statistical details and representative content moderation samples

We promote the indication of **representative content samples** of usual platform content (unflagged content sample/ appropriate content). These content samples should also be able to be viewed in the mock-up for the different roles included. Providing this context to the sample adding evidence about the reporting and the roles involved in the moderation process can help to understand the different digital environments and realities provided by the platforms in question.

Besides, we support the provision of information about the **sampling techniques** used for testing internal and external moderation decisions and developed representative content samples in line with Art 34 DSA and Art 37 DSA. A better understanding of **quality control of content moderation** can help to make provable and **empirical-based vetted researcher test and audit structures like the internal quality criteria** for online platforms. What statistical tools are used within such a process can substantially influence what is detected and what is out of scope.

Complexity and understandability of the terms and conditions, notice and action mechanisms, and out-of-court dispute settlement

We support the indication of standardised information about the understandability and complexity of the text that should guide users through the processes and rules within the content moderation process. This should help to create a user-friendly presentation of important rules and behavioural norms on the platform and should also indicate a measurement of comparability across service providers.

Role-action-time line

As mentioned above, we recommend the indication of a timeline that can visualize and log the content moderation actions, per content piece and per content. This can help to improve evidence about the moderation actions taken, by whom, and to provide information about which teams or individual moderators, or Trusted Flaggers were involved in the process. We therefore also propose to indicate the current content moderation status in such timelines. Providing insight into the timelines should also help to improve the quality of reporting for cases of suspension, demonetization, or other restrictions imposed on a user's account.

Therefore, we also support indicating not only the aggregated numbers of e.g., suspensions but rather to indicate the position within these timelines. This can help to answer also questions about, how long until a suspension is relieved from an account? Are there accounts which are permanently banned? How many accounts are not monetizing their content due to restrictions?

Findable Flags and User-Action-Timelines

We also propose to include information on when **individual flags were received per piece of content and when the moderation action was granted**. This could be illustrated as a **timeline** that maps out what moderation actions took place over the lifetime of a piece of content or account. Additionally, we would propose organizational and technical means that fulfil a '**find my flag'** function by using **unique IDs** to track flagging across the platform's notice takedown mechanism. Additionally providing a better general overview of the flagging behaviour across platforms – by users per Member State and trusted flaggers. An additional category might be addressing the public authorities and indications of Member States that make clear **which authority has requested what information or action from the platform in question in a higher-level descriptive way**. Also, the **unique ID** category **`Vetted Researcher´** might be insightful for data analysis according to Art 40 DSA.

Information about Submitted and not fully Submitted Notices

Providing this information can also help to better understand the ecosystem of moderation and the behaviour of the humans included in this complex interplay. If information about not fully submitted notices can also improve contextualisation in line with Art 16 DSA and provide the missing information about the flagging behaviour and could indicate obstacles in use early.

Machine-readability and processability

Translating various roles and processes to meaningful (shared) reporting structures

The Lumen database is using **various roles** within the content moderation process and transforms it into the structure of the database itself. These might e.g., be referred to as `entity_notice_roles_attributes´ in the Lumen database ('Lumen Database' [2012] 2023). Using unique IDs for reporting across the content moderation process allows tracking decisions and roles across reports and ideally across platforms. Indicating subcategories within the database used in line with **Art 17 DSA** could help to better understand the hierarchy of decisions taken within e.g., Member State team of content moderators.

CONCLUSION

Summarising, we can see that qualitative transparency reporting is a difficult art that has to understand the various addressees that have to be informed though it. Good transparency reporting also has to understand the content it is reporting about, its depth, design, mode of interaction and meaning. Different rules and regulations are guiding the way of compliant content moderation practises and accurate measurement and reporting, while also leaving platforms with room for interpretation and uncertainty about how those norms should be understood. Additionally, transparency reporting can't look back on a long history of reports yet and novel regulatory endeavours like the DSA show new needs for insight into the content moderation process as such. We also have to acknowledge the ever-changing challenges that content moderation, and its reporting has to deal with. Therefore, transparency reporting has to develop over time and address new issues and events while also understanding that new visualisation or details within reports might be needed. Transparency reports as such also have to be seen in the context of internal and external audits, advertising databases, the responsibility of Trusted Flaggers and the access grated to researchers. Only by analysing all these parts of fractioned insight thoroughly we can grasp a glimpse of inner workings and judge the proper protection and degree of standard upheld by platforms.

LIST OF RECOMENDATIONS

1. Different contexts for different roles and users
2. Different roles are involved in socio-technical processes
3. Unique IDs for content moderation roles
4. Reporting on the hierarchies of content moderation
5. Detailed information about the human support in content moderation
6. Example decisions and explanations
7. Reporting on relevant organisational and financial changes for VLOPS and VLOSES
8. Moderation times and moderation quality
9. Granular reporting for VLOPS and VLOSEs on legal and other capacities of content moderation teams
10. Role-Action Timeline for content moderation
11. Strengthening reporting on soft moderation
12. Mock-Up User and Moderation Interfaces for Reporting Purposes
13. Coherent counting and multiple indication approaches to reporting about content actioned
14. Reporting on Affected Content
15. Reporting per content type
16. Reporting on accuracy rates for automated tools per content category
17. Reporting in line with (perceived) systemic risk categories by design
18. Coherent counting and multiple indication approaches to reporting about content actioned
19. Comparing overblocking biases and incentives
20. Design of the notification, complaint handling, and dispute resolution systems
21. Systemic reporting on malicious attacks and moderation action
22. Agreed on accuracy rates for the use of automated tools
23. Provision of statistical details and representative content moderation samples
24. Complexity and understandability of the terms and conditions, notice and action mechanisms, and out-of-court dispute settlement
25. Role-action-time line
26. Findable Flags and User-Action-Timelines
27. Information about Submitted and not fully Submitted Notices
28. Translating various roles and processes to meaningful (shared) reporting structures

LITERATURE

Ahn, Soyun, Jeeyun (Sophia) Baik, and Clara Sol Krause. 2022. 'Splintering and Centralizing Platform Governance: How Facebook Adapted Its Content Moderation Practices to the Political and Legal Contexts in the United States, Germany, and South Korea'. *Information, Communication & Society* 0 (0): 1–20. https://doi.org/10.1080/1369118X.2022.2113817.

Alizadeh, Meysam, Fabrizio Gilardi, Emma Hoes, K. Jonathan Klüser, Maël Kubli, and Nahema Marchal. 2022. 'Content Moderation As a Political Issue: The Twitter Discourse Around Trump's Ban'. *Journal of Quantitative Description: Digital Media* 2 (October). https://doi.org/10.51685/jqd.2022.023.

'Api Documentation - DSA Transparency Database'. n.d. Accessed 17 July 2023. https://transparency.dsa.ec.europa.eu/page/api-documentation#statement-attributes.

Belli, Luca, Nicolo Zingales, and Yasmin Curzi. 2021. 'Glossary of Platform Law and Policy Terms'.

Bradshaw, S., and P. Howard. 2017. 'Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation'. *Computational Propaganda Research Project*. https://ora.ox.ac.uk/objects/uuid:cef7e8d9-27bf-4ea5-9fd6-855209b3e1f6.

Brignull, Harry. 2019. 'Dark Patterns'. Dark Patterns. 2019. https://www.darkpatterns.org/.

Buckley, Nicole, and Joseph S Schafer. 2022. '"Censorship-Free" Platforms: Evaluating Content Moderation Policies and Practices of Alternative Social Media' 4 (1).

Bundesgesetz Über Maßnahmen Zum Schutz Der Nutzer Auf Kommunikationsplattformen (Kommunikationsplattformen-Gesetz – KoPl-G – Communication Platforms Act). 2020. https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=20011415.

Chase, Peter H. 2019. 'The EU Code of Practice on Disinformation: The Difficulty of Regulating a Nebulous Problem'.

Christensen, Lars Thøger, and Joep Cornelissen. 2015. 'Organizational Transparency as Myth and Metaphor'. *European Journal of Social Theory* 18 (2): 132–49. https://doi.org/10.1177/1368431014555256.

Clune, Conor, and Emma McDaid. 2023. 'Content Moderation on Social Media: Constructing Accountability in the Digital Space'. *Accounting, Auditing & Accountability Journal* ahead-of-print (ahead-of-print). https://doi.org/10.1108/AAAJ-11-2022-6119.

Code of Practice on Disinformation. 2018.

Crawford, Kate, and Tarleton Gillespie. 2016. 'What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint'. *New Media & Society* 18 (3): 410–28. https://doi.org/10.1177/1461444814543163.

'Current Approaches to Terrorist and Violent Extremist Content among the Global Top 50 Online Content-Sharing Services'. 2020. OECD Digital Economy Papers 296. Vol. 296. OECD

Digital Economy Papers. https://doi.org/10.1787/68058b95-en.

Dinar, Christina, and Lena Hinrichs. 2022. '(Niche) Platforms as Experimental Spaces for Content Moderation - Mapping Small, Medium and Niche Platforms Online'.

Douek, Evelyn. 2022. 'Content Moderation as Systems Thinking'. *Harvard Law Review* 136: 526.

Echikson, William, and Olivia Knodt. 2018. 'Germany s NetzDG: A Key Test for Combatting Online Hate'. SSRN Scholarly Paper ID 3300636. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=3300636.

Einwiller, Sabine A., and Sora Kim. 2020. 'How Online Content Providers Moderate User-Generated Content to Prevent Harmful Online Communication: An Analysis of Policies and Their Implementation'. *Policy & Internet* 12 (2): 184–206. https://doi.org/10.1002/poi3.239.

'European Economic Area (EEA) - Code of Practice on Disinformation'. 2023. TikTok. 7 February 2023. https://www.tiktok.com/transparency/en/copd-eu/.

European Parliament and the Council, and Council of Europe. 2021. 'Directive on Improving Working Conditions in Platform Work'.

Ferrara, Emilio, Herbert Chang, Emily Chen, Goran Muric, and Jaimin Patel. 2020. 'Characterizing Social Media Manipulation in the 2020 U.S. Presidential Election'. *First Monday*, October. https://doi.org/10.5210/fm.v25i11.11431.

Fiala, Lenka, and Martin Husovec. 2022. 'Using Experimental Evidence to Improve Delegated Enforcement'. *International Review of Law and Economics* 71 (September): 106079. https://doi.org/10.1016/j.irle.2022.106079.

Fischer, Caroline, and Sascha Kraus. 2022. 'Digitale Transparenz'. In *Handbuch Digitalisierung in Staat und Verwaltung*, 1–13. Springer. https://doi.org/10.1007/978-3-658-23669-4_14-2.

Flyverbom, Mikkel. 2015. 'Sunlight in Cyberspace? On Transparency as a Form of Ordering'. *European Journal of Social Theory* 18 (2): 168–84. https://doi.org/10.1177/1368431014555258.

———. 2016a. 'Digital Age| Transparency: Mediation and the Management of Visibilities'. *International Journal of Communication* 10 (0): 13.

———. 2016b. 'Disclosing and Concealing: Internet Governance, Information Control and the Management of Visibility'. *Internet Policy Review* 5 (3). https://policyreview.info/articles/analysis/disclosing-and-concealing-internet-governance-information-control-and-management.

Flyverbom, Mikkel, Ronald Deibert, and Dirk Matten. 2019. 'The Governance of Digital Technology, Big Data, and the Internet: New Roles and Responsibilities for Business'. *Business & Society* 58 (1): 3–19. https://doi.org/10.1177/0007650317727540.

Flyverbom, Mikkel, Anders Koed Madsen, and Andreas Rasche. 2017. 'Big Data as Governmentality in International Development: Digital Traces, Algorithms, and Altered Visibilities'. *The Information Society* 33 (1): 35–42.

https://doi.org/10.1080/01972243.2016.1248611.

Flyverbom, Mikkel, and John Murray. 2018. 'Datastructuring—Organizing and Curating Digital Traces into Action'. *Big Data & Society* 5 (2): 205395171879911. https://doi.org/10.1177/2053951718799114.

Gesetz Zur Verbesserung Der Rechtsdurchsetzung in Sozialen Netzwerken – Network Enforcement Act. 2017a. *BGBl I S 2252*. https://www.bmj.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf;jsessionid=2A0E86949FA3381D1C170D3FEFB4C20B.1_cid324?__blob=publicationFile&v=2.

———. 2017b. *BGBl I S 2252*. https://www.bmj.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf;jsessionid=2A0E86949FA3381D1C170D3FEFB4C20B.1_cid324?__blob=publicationFile&v=2.

Giansiracusa, Noah. 2021. 'How Facebook Hides How Terrible It Is With Hate Speech'. *Wired*, 2021. https://www.wired.com/story/facebooks-deceptive-math-when-it-comes-to-hate-speech/.

'GIFCT'. n.d. GIFCT. Accessed 19 August 2023. https://gifct.org/.

Golia, Angelo Jr, Matthias C. Kettemann, and Raffaela Kunz. 2021. 'International Law and the Internet'. *Zeitschrift Für Ausländisches Öffentliches Recht Und Völkerrecht / Heidelberg Journal of International Law* 81 (3): 597–600. https://doi.org/10.17104/0044-2348-2021-3-597.

Gollatz, Kirsten, Martin Riedl, and Jens Pohlmann. 2018. 'Removals of Online Hate Speech in Numbers', August. https://doi.org/10.5281/zenodo.1342324.

Gollatz, Kirsten/Riedl. 2018. 'Removals Of Online Hate Speech In Numbers', August. https://doi.org/10.5281/ZENODO.1342324.

Gorwa, Robert. 2021. 'Elections, Institutions, and the Regulatory Politics of Platform Governance: The Case of the German NetzDG'. *Telecommunications Policy*, Norm entrepreneurship in Internet Governance, 45 (6): 102145. https://doi.org/10.1016/j.telpol.2021.102145.

Heldt, Amelie. 2019. 'Reading between the Lines and the Numbers: An Analysis of the First NetzDG Reports'. *Internet Policy Review*. https://doi.org/10.14763/2019.2.1398.

Heldt, Amélie. 2019. 'Reading between the Lines and the Numbers: An Analysis of the First NetzDG Reports'. *Internet Policy Review* 8 (2). https://policyreview.info/articles/analysis/reading-between-lines-and-numbers-analysis-first-netzdg-reports.

Heldt, Amélie Pia. 2018. 'Reading Between the Lines and the Numbers: An Analysis of the First NetzDG Reports'. SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=3413677.

Hemmert-Halswick, Maximilian. 2021. 'Lessons Learned from the First Years with the NetzDG'. In , 415–32. Nomos Verlagsgesellschaft mbH & Co. KG. https://doi.org/10.5771/9783748929789-415.

Hoffman, Stephen P. 2010. ‘An Illustration of Hashing and Its Effect on Illegal File Content in the Digital Age’. SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=1564980.

Hovyadinov, Sergei. 2019. ‘Toward a More Meaningful Transparency: Examining Twitter, Google, and Facebook’s Transparency Reporting and Removal Practices in Russia’. SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.3535671.

Institute for Strategic Dialog. 2023. ‘Deny, Deceive, Delay: Exposing New Trends in Climate Mis- and Disinformation at COP27’. 2. https://www.isdglobal.org/wp-content/uploads/2023/01/Deny-Deceive-Delay-Vol.-2.pdf.

Juneja, Prerna, Deepika Rama Subramanian, and Tanushree Mitra. 2020. ‘Through the Looking Glass: Study of Transparency in Reddit’s Moderation Practices’. *Proceedings of the ACM on Human-Computer Interaction* 4 (GROUP): 17:1-17:35. https://doi.org/10.1145/3375197.

Kalker, Ton, Jaap Haitsma, and Job C. Oostveen. 2001. ‘Issues with Digital Watermarking and Perceptual Hashing’. In *Multimedia Systems and Applications IV*, 4518:189–97. SPIE. https://doi.org/10.1117/12.448203.

Kaminski, Margot E. 2020. ‘Understanding Transparency in Algorithmic Accountability’. SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=3622657.

Katsaros, Matthew, Jisu Kim, and Tom Tyler. 2023. ‘Online Content Moderation: Does Justice Need a Human Face?’ *International Journal of Human–Computer Interaction* 0 (0): 1–12. https://doi.org/10.1080/10447318.2023.2210879.

Kaye, David. 2018. ‘Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression’. A/HRC/38/35. Geneva, Switzerland: United Nations. https://documents-dds-ny.un.org/doc/UNDOC/GEN/G18/096/72/PDF/G1809672.pdf?OpenElement.

Keller, Daphne. 2019. ‘Who Do You Sue? State and Platform Hybrid Power Over Online Speech’. *Hoover Institution Essay*, 40.

———. 2023. ‘Rushing to Launch the EU’s Platform Database Experiment’. 16 July 2023. https://cyberlaw.stanford.edu/blog/2023/07/rushing-launch-eus-platform-database-experiment.

Keller, Daphne, and Paddy Leerssen. 2019. ‘Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation’. SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=3504930.

Kettemann, Matthias C, and Marie-Therese Sekwenz. 2022. *Pandemocracy in Europe: Power, Parliaments and People in Times of COVID-19*. Edited by Matthias C Kettemann and Konrad Lachmayer. Hart Publishing. https://doi.org/10.5040/9781509946396.

Kettemann, Matthias C., and Anna Sophia Tiedeke. 2020. ‘Back up: Can Users Sue Platforms to Reinstate Deleted Content?’ *Internet Policy Review* 9 (2). https://policyreview.info/articles/analysis/back-can-users-sue-platforms-reinstate-deleted-

content.

Kettemann, Matthias, Felicitas Rachinger, and Marie-Therese Sekwenz. 2022. ‘Deplatforming’. In , 78–89. https://doi.org/10.5771/9783214164935-78.

Kübler, Johanne, Marie-Theres Sekwenz, Felicitas Rachinger, Anna König, Rita Gsenger, Eliška Pírková, Ben Wagner, Matthias C Kettemann, Michael Krennerich, and Carolina Ferro. 2023. ‘The 2021 German Federal Election on Social Media: Analysing Electoral Risks Created by Twitter and Facebook’. In . Hawaii.

Lakens, Daniël, and Lisa M. DeBruine. 2021. ‘Improving Transparency, Falsifiability, and Rigor by Making Hypothesis Tests Machine-Readable’. *Advances in Methods and Practices in Psychological Science* 4 (2): 251524592097094. https://doi.org/10.1177/2515245920970949.

Ling, Chen, Krishna Gummadi, and Savvas Zannettou. 2022. *‘Learn the Facts About COVID-19’: Analyzing the Use of Warning Labels on TikTok Videos*.

Llanos, J. 2021a. ‘Transparency Reporting: Considerations for the Review of the OECD Privacy Guidelines’. Report 309. *(OECD Digital Economy Papers 309). OECD Publishing: Paris, Fance.* Paris, Fance: OECD Publishing. https://doi.org/10.1787/e90c11b6-en.

———. 2021b. ‘Transparency Reporting on Terrorist and Violent Extremist Content Online: An Update on the Global Top 50 Content Sharing Services’. Report 313. *(OECD Digital Economy Papers 313). OECD Publishing, Paris: Paris, Fance.* Paris, Fance: OECD Publishing, Paris. https://doi.org/10.1787/8af4ab29-en.

LOI N° 2020-766 Du 24 Juin 2020 Visant à Lutter Contre Les Contenus Haineux Sur Internet (1). 2020. *2020-766*.

‘Lumen Database’. (2012) 2023. Ruby. Berkman Klein Center for Internet & Society. https://github.com/berkmancenter/lumendatabase.

Ma, Renkai, and Yubo Kou. 2022. ‘“I’m Not Sure What Difference Is between Their Content and Mine, Other than the Person Itself”: A Study of Fairness Perception of Content Moderation on YouTube’. *Proceedings of the ACM on Human-Computer Interaction* 6 (CSCW2): 425:1-425:28. https://doi.org/10.1145/3555150.

MacCarthy, Mark. 2019. ‘An Examination of the Algorithmic Accountability Act of 2019’. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3615731.

———. 2020a. ‘Transparency Requirements for Digital Social Media Platforms’.

———. 2020b. ‘Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry’. SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.3615726.

———. 2022. ‘Transparency Recommendations for Regulatory Regimes of Digital Platforms’. Report. Centre for International Governance Innovation. https://apo.org.au/node/316845.

Mancuso (CELA), Shiho. 2020. ‘The Global Internet Forum to Counter Terrorism’, July. https://policycommons.net/artifacts/3855072/the-global-internet-forum-to-counter-terrorism/4661027/.

Michener, Gregory. 2019. ‘Gauging the Impact of Transparency Policies’. *Public Administration Review* 79 (1): 136–39. https://doi.org/10.1111/puar.13011.

Miller, Carl, David Weir, Shaun Ring, Oliver Marsh, Chris Inskip, and Nestor Prieto Chavana. 2023. ‘Antisemitism on Twitter Before and After Elon Musk’s Acquisition’.

Myers West, Sarah. 2018. ‘Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms’. *New Media & Society* 20 (11): 4366–83. https://doi.org/10.1177/1461444818773059.

NetzDGÄndG. 2021. https://www.gesetze-im-internet.de/netzdg/index.html.

Niu, Yanhua, Shuai Gao, and Hongke Zhang. 2023. ‘A Trustworthy Content Moderation Scheme Based on Permissioned Blockchain’. In *Emerging Networking Architecture and Technologies*, edited by Wei Quan, 131–45. Communications in Computer and Information Science. Singapore: Springer Nature. https://doi.org/10.1007/978-981-19-9697-9_11.

Nunziato, Dawn Carla. 2023. ‘The Digital Services Act and the Brussels Effect on Platform Content Moderation’. SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=4425793.

OECD. 2021. ‘Transparency Reporting on Terrorist and Violent Extremist Content Online: An Update on the Global Top 50 Content Sharing Services’. Paris: OECD. https://doi.org/10.1787/8af4ab29-en.

———. 2022. ‘Transparency Reporting on Terrorist and Violent Extremist Content Online 2022’. Paris: OECD. https://doi.org/10.1787/a1621fc3-en.

Pan, Christina A., Sahil Yakhmi, Tara P. Iyer, Evan Strasnick, Amy X. Zhang, and Michael S. Bernstein. 2022. ‘Comparing the Perceived Legitimacy of Content Moderation Processes: Contractors, Algorithms, Expert Panels, and Digital Juries’. *Proceedings of the ACM on Human-Computer Interaction* 6 (CSCW1): 82:1-82:31. https://doi.org/10.1145/3512929.

Park, Jieum. 2020. ‘The Public-Private Partnerships’ Impact on Transparency and Effectiveness in the EU Internet Content Regulation: The Case of “Network Enforcement Act (NetzDG)” in Germany’. Master Thesis, Potsdam: Universitätsverlag Potsdam. https://doi.org/10.25932/publishup-48718.

Radu, Roxana, Matthias C. Kettemann, Trisha Meyer, and Jamal Shahin. 2021. ‘Normfare: Norm Entrepreneurship in Internet Governance’. *Telecommunications Policy*, Norm entrepreneurship in Internet Governance, 45 (6): 102148. https://doi.org/10.1016/j.telpol.2021.102148.

Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on Addressing the Dissemination of Terrorist Content Online (Text with EEA Relevance). 2021. *OJ L*. Vol. 172. http://data.europa.eu/eli/reg/2021/784/oj/eng.

Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and Amending Directive 2000/31/EC (Digital Services Act) (Text with EEA Relevance). 2022. *OJ L*. Vol. 277. http://data.europa.eu/eli/reg/2022/2065/oj/eng.

‘Removals under the Network Enforcement Law – Google Transparency Report’. n.d. Accessed 23 July 2023. https://transparencyreport.google.com/netzdg/youtube?hl=en.

Reuters. 2019. ‘Germany Fines Facebook for Under-Reporting Complaints’, 2 July 2019, sec. Media Industry. https://www.reuters.com/article/us-facebook-germany-fine-idUSKCN1TX1IC.

Sander, Barrie. 2019. ‘Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation’. *Fordham International Law Journal* 43: 939.

‘Santa Clara Principles on Transparency and Accountability in Content Moderation’. n.d. Santa Clara Principles. Accessed 2 June 2023a. https://santaclaraprinciples.org/images/santa-clara-OG.png.

‘———’. n.d. Santa Clara Principles. Accessed 2 June 2023b. https://santaclaraprinciples.org/images/santa-clara-OG.png.

‘———’. n.d. Santa Clara Principles. Accessed 2 June 2023c. https://santaclaraprinciples.org/images/santa-clara-OG.png.

Schulz, Wolfgang. 2022a. ‘Regulating Intermediaries to Protect Personality Rights Online—The Case of the German NetzDG’. In *Personality and Data Protection Rights on the Internet: Brazilian and German Approaches*, edited by Marion Albers and Ingo Wolfgang Sarlet, 289–307. Ius Gentium: Comparative Perspectives on Law and Justice. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-90331-2_12.

———. 2022b. ‘Regulating Intermediaries to Protect Personality Rights Online—The Case of the German NetzDG’. In *Personality and Data Protection Rights on the Internet: Brazilian and German Approaches*, edited by Marion Albers and Ingo Wolfgang Sarlet, 289–307. Ius Gentium: Comparative Perspectives on Law and Justice. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-90331-2_12.

Schulz, Wolfgang, Matthias C. Kettemann, and Amélie P. Heldt. 2019. ‘Probleme und Potenziale des NetzDG - ein Reader mit fünf HBI-Expertisen’. *Arbeitspapiere des Hans-Bredow-Instituts*. https://doi.org/10.21241/SSOAR.71727.

Shahid, Farhana, and Aditya Vashistha. 2023. ‘Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony?’ In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18. CHI ’23. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3544548.3581538.

Solomun, Sonja, Maryna Polataiko, and Helen A Hayes. 2020. ‘Platform Responsability and Refulation in Canada: Consinderations on Transparency, Legislative Clarity, and Deisgn’ 34.

Son, Heui-Su, Sung-Woo Byun, and Seok-Pil Lee. 2020. ‘A Robust Audio Fingerprinting Using a New Hashing Method’. *IEEE Access* 8: 172343–51. https://doi.org/10.1109/ACCESS.2020.3024951.

Song, Jean Y., Sangwook Lee, Jisoo Lee, Mina Kim, and Juho Kim. 2023. ‘ModSandbox:

Facilitating Online Community Moderation Through Error Prediction and Improvement of Automated Rules'. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–20. CHI '23. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3544548.3581057.

Sontheimer, Lukas, Johannes Schäfer, and Thomas Mandl. 2022. 'Enabling Informational Autonomy through Explanation of Content Moderation: UI Design for Hate Speech Detection'. https://doi.org/10.18420/muc2022-mci-ws12-260.

Steinebach, Martin, Huajian Liu, and York Yannikos. 2012. 'ForBild: Efficient Robust Image Hashing'. In *Media Watermarking, Security, and Forensics 2012*, 8303:195–202. SPIE. https://doi.org/10.1117/12.907856.

Thach, Hibby, Samuel Mayworm, Daniel Delmonaco, and Oliver Haimson. 2022. '(In)Visible Moderation: A Digital Ethnography of Marginalized Users and Content Moderation on Twitch and Reddit'. *New Media & Society*, July, 14614448221109804. https://doi.org/10.1177/14614448221109804.

The EU Code of Conduct on Countering Illegal Hate Speech Online. 2019. *2016*. https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en.

'The Strengthened Code of Practice on Disinformation'. 2022. Centre for Media Pluralism and Freedom. 2022. https://cmpf.eui.eu/event/strengthened-code-of-practice-on-disinformation/.

Tiedeke, Anna-Sophia, Matthias C. Kettemann, Felicitas Rachinger, Marie-Theres Sekwenz, and Ben Wagner. 2020. 'What Can Be Said Online in Germany and Austria? A Legal and Terms of Service Taxonomy'. SSRN Scholarly Paper ID 3735932. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.3735932.

Tworek, Heidi, and Paddy Leerssen. 2019. 'An Analysis of Germany's NetzDG Law'.

———. n.d. 'An Analysis of Germany's NetzDG Law', 11.

Urban, Jennifer, and Laura Quilter. 2006. 'Efficient Process or Chilling Effects - Takedown Notices under Section 512 of the Digital Millennium Copyright Act'. *Santa Clara High Technology Law Journal* 22 (4): 621.

Urman, Aleksandra, and Mykola Makhortykh. 2023. 'How Transparent Are Transparency Reports? Comparative Analysis of Transparency Reporting across Online Platforms'. *Telecommunications Policy* 47 (3): 102477. https://doi.org/10.1016/j.telpol.2022.102477.

Wagner, Ben. 2020. 'Regulating Transparency? Facebook, Twitter and the German Network Enforcement Act.', 261–71. https://doi.org/10.1145/3351095.3372856.

Wagner, Ben, Matthias C Kettemann, Anna-Sophia Tiedeke, Marie-Therese Sekwenz, and Felicitas Rachinger. 2021. 'Forthcoming. Blurring Legal Boundaries. Recoding Interpretations of Law and Terms of Service in Online Content Governance'.

Wagner, Ben, Johanne Kübler, Eliška Pírková, Rita Gsenger, and Carolina Ferro. n.d.

‘REIMAGINING CONTENT MODERATION AND SAFEGUARDING FUNDAMENTAL RIGHTS’.

Wagner, Ben, Krisztina Rozgonyi, Marie-Therese Sekwenz, Jennifer Cobbe, and Jatinder Singh. 2020a. ‘Regulating Transparency? Facebook, Twitter and the German Network Enforcement Act’. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 261–71. FAT* ’20. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372856.

———. 2020b. ‘Regulating Transparency? Facebook, Twitter and the German Network Enforcement Act’. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 261–71. FAT* ’20. Barcelona, Spain: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372856.

———. 2020c. ‘Regulating Transparency? Facebook, Twitter and the German Network Enforcement Act’. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 261–71. FAT* ’20. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372856.

Wang, Wenxuan, Jingyuan Huang, Chang Chen, Jiazhen Gu, Jianping Zhang, Weibin Wu, Pinjia He, and Michael Lyu. 2023. ‘Validating Multimedia Content Moderation Software via Semantic Fusion’. arXiv. https://doi.org/10.48550/arXiv.2305.13623.

Westlake, Bryce, Martin Bouchard, and Richard Frank. 2012. ‘Comparing Methods for Detecting Child Exploitation Content Online’. In *2012 European Intelligence and Security Informatics Conference*, 156–63. https://doi.org/10.1109/EISIC.2012.25.

Yannikos, York, Nadeem Ashraf, Martin Steinebach, and Christian Winter. 2013. ‘Automating Video File Carving and Content Identification’. In *Advances in Digital Forensics IX*, edited by Gilbert Peterson and Sujeet Shenoi, 195–212. IFIP Advances in Information and Communication Technology. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-41148-9_14.

Ye, Meng, Karan Sikka, Katherine Atwell, Sabit Hassan, Ajay Divakaran, and Malihe Alikhani. 2023. ‘Multilingual Content Moderation: A Case Study on Reddit’. arXiv. https://doi.org/10.48550/arXiv.2302.09618.

Zannettou, Savvas. 2021. ‘“I Won the Election!”: An Empirical Analysis of Soft Moderation Interventions on Twitter’. *Proceedings of the International AAAI Conference on Web and Social Media* 15 (May): 865–76. https://doi.org/10.1609/icwsm.v15i1.18110.

Zhao, Andy, and Zhaodi Chen. 2023. ‘Let’s Report Our Rivals: How Chinese Fandoms Game Content Moderation to Restrain Opposing Voices’. *Journal of Quantitative Description: Digital Media* 3 (April). https://doi.org/10.51685/jqd.2023.006.

Zipursky, Rebecca. 2019. ‘Nuts About NETZ: The Network Enforcement Act and Freedom of Expression’. *Fordham International Law Journal* 42 (4): 1325.

Zornetta, Alessia, and Ilka Pohland. 2022. ‘Legal and Technical Trade-Offs in the Content Moderation of Terrorist Live-Streaming’. *International Journal of Law and Information Technology* 30 (3): 302–20. https://doi.org/10.1093/ijlit/eaac020.

ANNEXES

1.1	**The Digital Services Act (DSA)** (transparency-norm classes)			**Transparency Reporting Category (TRC)**		
	Transparency Reporting Quality Categories (quality category**)**	**intermediary services**	**Hosting**	**Online Platform**	**Online Platform with Means for Distance Contract**	**VLOP/VLSE**
	Potentially Relevant Legal Norms					
1.1.2.1.	**Contextualization** quality category**:**					
1.1.2.1.1.	**Transparency Reporting Times (TNTs)**					
1.1.2.2.	**Comparability** quality category					
1.1.2.2.1.	Transparency Reporting Times (TNTs)					
1.1.3.	**Vulnerability to Manipulation** quality category					
1.1.3.1.	Transparency Reporting Times (TNTs)					
1.1.4.	**Provable** quality category					
1.1.4.1.	Transparency Reporting Times (TNTs)					
1.1.5.	**Machine-Readability and Processability** quality category					
1.1.5.1.	**Transparency Reporting Times (TNTs)**					

1.2	**Regulation (EU) 2021/784 on Terrorist Content Online (**transparency-norm classes**)**			**Transparency Reporting Category (TRC)**		
	Transparency Reporting Quality Categories (quality category**)**	**intermediary services**	**Hosting**	**Online Platform**	**Online Platform with Means for Distance Contract**	**VLOP/VLSE**
1.2.2.1.	**Contextualization** quality category					
1.2.2.1.1.	Transparency Reporting Times (TNTs)					
1.2.2.2.	**Comparability** quality category					
1.2.2.2.1.	Transparency Reporting Times (TNTs)					
1.2.3.	**Vulnerability to Manipulation** quality category					
1.2.3.1.	Transparency Reporting Times (TNTs)					
1.2.4.	**Provable** quality category					
1.2.4.1.	Transparency Reporting Times (TNTs)					
1.2.5.	**Machine-Readability and Processability** quality category					
1.2.5.1.	**Transparency Reporting Times (TNTs)**					

1.3	**The German Network Enforcement Act (NetzDG)** (transparency-norm classes)			**Transparency Reporting Category (TRC)**		
	Transparency Reporting Quality Categories (quality category**)**	**intermediary services**	**Hosting**	**Online Platform**	**Online Platform with Means for Distance Contract**	**VLOP/VLSE** platforms that receive more than 100 complaints about unlawful content per calendar year have transparency reporting duties (§ 2 (1))
1.3.2.1.	**Contextualization** quality category					
1.3.2.1.1.	Transparency Reporting Times (TNTs)					
1.3.2.2.	**Comparability** quality category					
1.3.2.2.1.	Transparency Reporting Times (TNTs)					
1.3.3.	**Vulnerability to Manipulation** quality category					
1.3.3.1.	Transparency Reporting Times (TNTs)					
1.3.4.	**Provable** quality category					
1.3.4.1.	Transparency Reporting Times (TNTs)					

1.3.5.	**Machine-Readability and Processability** quality category					
1.3.5.1.	**Transparency Reporting Times (TNTs)**					

2.1	**OECD Terrorist and Violent Extremist Content (TVEC)** (transparency-norm classes)			**Transparency Reporting Category (TRC)**		
	Transparency Reporting Quality Categories (quality category)	**intermediary services**	**Hosting**	**Online Platform online content-sharing services**	**Online Platform with Means for Distance Contract**	**VLOP/VLSE**
2.1.2.1.	**Contextualization** quality category					
2.1.2.1.1.	Transparency Reporting Times (TNTs)					
2.1.2.2.	**Comparability** quality category					
2.1.2.2.1.	Transparency Reporting Times (TNTs)					
2.1.3.	**Vulnerability to Manipulation** quality category					
2.1.3.1.	Transparency Reporting Times (TNTs)					
2.1.4.	**Provable** quality category					
2.1.4.1.	Transparency Reporting Times (TNTs)					
2.1.5.	**Machine-Readability and Processability** quality category					

2.1.5.1.	**Transparency Reporting Times (TNTs)**					

2.2	**The Santa Clara Principles on Transparency and Accountability (**transparency-norm classes)			**Transparency Reporting Category (TRC)**		
	Transparency Reporting Quality Categories (quality category**)**	**intermediary services**	**Hosting**	**Online Platform**	**Online Platform with Means for Distance Contract**	**VLOP/VLSE**
2.2.2.1.	**Contextualization** quality category					
2.2.2.1.1.	Transparency Reporting Times (TNTs)					
2.2.2.2.	**Comparability** quality category					
2.2.2.2.1.	Transparency Reporting Times (TNTs)					
2.2.3.	**Vulnerability to Manipulation** quality category					
2.2.3.1.	Transparency Reporting Times (TNTs)					
2.2.4.	**Provable** quality category					
2.2.4.1.	Transparency Reporting Times (TNTs)					
2.2.5.	**Machine-Readability and Processability** quality category					

2.2.5.1.	**Transparency Reporting Times (TNTs)**					

3.1.	**Relevant codes of conduct with transparency measures, such as the Code of Practice on Disinformation (**transparency-norm classes)			**Transparency Reporting Category (TRC)**		
	Transparency Reporting Quality Categories (quality category**)**	**intermediary services**	**Hosting**	**Online Platform**	**Online Platform with Means for Distance Contract**	**VLOP/VLSE**
3.1.2.1.	**Contextualization** quality category					
3.1.2.1.1.	Transparency Reporting Times (TNTs)					
3.1.2.2.	**Comparability** quality category					
3.1.2.2.1.	Transparency Reporting Times (TNTs)					
3.1.3.	**Vulnerability to Manipulation** quality category					
3.1.3.1.	Transparency Reporting Times (TNTs)					
3.1.4.	**Provable** quality category					
3.1.4.1.	Transparency Reporting Times (TNTs)					
3.1.5.	**Machine-Readability and Processability** quality category					

3.1.5.1.	**Transparency Reporting Times (TNTs)**					

	Transparency Norm Classes	**Contextualization Transparency Reporting Category**	Benefits Contextualization Transparency Reporting Category	Drawbacks Contextualization Transparency Reporting Category	**Comparability Transparency Reporting Category**	Benefits Comparability Transparency Reporting Category
1	**Obligatory Legal Transparency Norm Class**					
1.1	The Digital Services Act (DSA)					
1.2	Regulation (EU) 2021/784 on Terrorist Content Online					
1.3	The German Network Enforcement Act (NetzDG)					
2	**Voluntary Reporting Transparency Norm Class**					
2.1	OECD Terrorist and Violent Extremist Content (TVEC)					
2.2	The Santa Clara Principles on Transparency and Accountability Around Content Moderation, and associated implementation toolkits for advocates, companies					
3	**Code of Conduct Transparency Norm Class**					
3.1	Relevant codes of conduct with transparency measures, such as the Code of Practice on Disinformation					

Drawbacks Comparability Transparency Reporting Category	**Vulnerability to Manipulation Transparency Reporting Category**	Benefits Vulnerability to Manipulation Transparency Reporting	Drawbacks Vulnerability to Manipulation Transparency Reporting	**Provability Transparency Reporting Category**	Benefits Provability Transparency Reporting Category	Drawbacks Provability Transparency Reporting Category	**Machine-Readability and Processability Transparency Reporting Category**

Benefits Machine-Readability and Processability Transparency Reporting	**Drawbacks Machine-Readability and Processability Transparency Reporting Category**	Benefits Machine-Readability and Processability Transparency Reporting Category	Benefits Drawbacks Machine-Readability and Processability Transparency Reporting Category				