How transparent are transparency reports? Comparative analysis of transparency reporting across online platforms

Urman, Aleksandra ; Makhortykh, Mykola

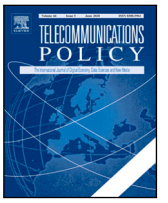
DOI: https://doi.org/10.1016/j.telpol.2022.102477

Telecommunications Policy 47 (2023) 102477

Contents lists available at ScienceDirect

Telecommunications Policy

journal homepage: www.elsevier.com/locate/telpol



How transparent are transparency reports? Comparative analysis of transparency reporting across online platforms

Aleksandra Urman [a,*], Mykola Makhortykh [b]

[a] Social Computing Group, Department of Informatics, University of Zurich, Zurich, Switzerland
[b] Institute of Communication and Media Studies, University of Bern, Bern, Switzerland

ARTICLE INFO

Keywords:
Transparency reports
Transparency
Online platforms
Santa Clara Principles
Compliance
Comparative analysis
Regulation
Digital Services Act

ABSTRACT

Over the last decade, transparency reports have been adopted by most large information technology companies. These reports provide important information on the requests tech companies receive from state actors around the world and the ways they respond to these requests, including what content the companies remove from platforms they own. In theory, such reports shall make inner workings of companies more transparent, in particular with respect to their collaboration with state actors. They shall also allow users and external entities (e.g., researchers or watchdogs) to assess to what extent companies adhere to their own policies on user privacy and content moderation as well as to the principles formulated by global entities that advocate for the freedom of expression and privacy online such as the Global Network Initiative or Santa Clara Principles. However, whether the current state of transparency reports actually is conducive to meaningful transparency remains an open question. In this paper, we aim to address this through a critical comparative analysis of transparency reports using Santa Clara Principles 2.0 (SCP 2.0) as the main analytical framework. Specifically, we aim to make three contributions: first, we conduct a comparative analysis of the types of data disclosed by major tech companies and social media platforms in their transparency reports. The companies and platforms analyzed include Google (incl. YouTube), Microsoft (incl. its subsidiaries Github and LinkedIn), Apple, Meta (prev. Facebook), TikTok, Twitter, Snapchat, Pinterest, Reddit and Amazon (incl. subsidiary Twitch). Second, we evaluate to what degree the released information complies with SCP 2.0 and how it aligns with different purposes of transparency. Finally, we outline recommendations that could improve the level of transparency within the reports and beyond, and contextualize our recommendations with regard to the Digital Services Act (DSA) that received the final approval of the European Council in October 2022.

1. Introduction

Transparency is one of key concepts that guide the debate on online platform governance and its sustainability. While multiple conceptualizations of transparency exist (Albu & Flyverbom, 2019; Gorwa & Garton Ash, 2020; Heald, 2006), it can be broadly defined as the practice of providing internal information "on matters of public concern" (Cotterrell, 1999, p. 414) by the companies owning the respective platforms to the external audience. While the reliance on transparency as an accountability mechanism has its own flaws (e.g., because of its temporal/technical limitations or focus on neoliberal forms of agency Ananny & Crawford, 2018), the ability to access information about platform functionalities and policies and their enforcement remains a key prerequisite for making the public, governments, and other stakeholders able to assess platforms' performance.

* Correspondence to: Andreasstrasse 15, 8050 Zurich, Switzerland.
E-mail address: urman@ifi.uzh.ch (A. Urman).

https://doi.org/10.1016/j.telpol.2022.102477
Received 21 April 2022; Received in revised form 30 October 2022; Accepted 31 October 2022
Available online 6 January 2023

Transparency can deal with different aspects of how platforms function, ranging from their business models (Wu, 2016) to the ways user data are processed (Sneed, 2020) and platform policies are enforced (Makhortykh, Urman, Münch et al., 2022). However, it is particularly important in the context of the platforms' societal impact. This is the case with platform content moderation, namely the set of policies and functionalities determining what forms of content are unacceptable within the platform and how they are dealt with Gillespie et al. (2020). The importance of making content moderation transparent is attributed not only to its notorious obscurity (Gorwa et al., 2020; Makhortykh, Urman, Münch et al., 2022), but also to the increasing use of platforms for political communication in both democratic and authoritarian contexts (O'Regan & Li, 2019; Stier et al., 2018; Urman, 2020; Urman et al., 2021). Under these circumstances, platform content moderation has potential for both preventing platform abuse (e.g., by removing extremist content Conway, 2020) and facilitating it (e.g., by requesting platforms to remove content criticizing the authorities Sivetc & Wijermars, 2021).

In fact, discussions about online platform transparency are tightly connected to the debates on the liability of online platforms with respect to content moderation. MacKinnon et al. (2015) identified three main types of regulatory approaches to online platforms in this context: broad immunity, conditional liability, strict liability. Broad immunity characterizes the current approach in the US, where most major online platforms were founded and have their headquarters. There, under the 230 Section of the US telecommunication law, online intermediaries are not liable for what their users post online — in other words, they are not considered "publishers" of their users' content in legal sense (Gillespie, 2018). This makes online platforms distinct from traditional publishers such as "old" media in terms of the way they are governed. Further, under the "broad immunity" provided by the 230 Section, the platforms are not liable for the content posted by the users even when platforms actively "police" user activities - e.g., through content moderation. Online platforms in the US thus have the right to remove user content but not the responsibility to do it at all or according to certain standards. It is worth noting that there are some important exceptions to this such as those related to the material that is illegal to distribute on federal level such as in copyright infringement cases (under Digital Millennium Copyright Act (DMCA)) or when it comes to the material violating federal and state sex trafficking laws (2018 amendment of 230 Section under FOSTA-SESTA).

The conditional liability approach, on the other hand, means that the platforms are not liable for the user content as long as they have no actual knowledge of illicit content and did not produce it themselves. They should remove the content upon the requests of the state or the courts. This approach is prevalent in Europe and most South American countries (MacKinnon et al., 2015). The strict liability approach adopted in China and many countries of the Middle East requires that online platforms proactively remove illicit content. This approach often implies close cooperation between the authorities and the platforms and results in platforms becoming part of authoritarian governments' online censorship practices (Gillespie, 2018). Finally, in some countries – e.g., some in sub-Saharan Africa, – there are currently no laws regulating online platforms at all.

It is evident that the liability of online platforms is regulated differently across the globe, and it is unlikely and perhaps even undesirable that the universal liability standards will be developed and applied to platforms. In light of this, as well as constant changes in the regulation of online platforms attributed to new forms of legislation such as the European Digital Services Act (see below), transparency becomes all the more important. It serves to inform the policy decisions regarding online platforms, helps individuals to make informed decisions about the use of specific platforms and the public to put pressure on the platforms (MacCarthy, 2020). Transparency thus serves different purposes, which can vary depending on the type of regulatory approach to platforms and, more broadly, political regime in a given context. For example, user empowerment is especially important under the conditions of strict platform liability, especially in autocracies, because greater transparency can provide users crucial information about the scope of the platforms' cooperation with the governments, including in ways that aid censorship and compromise users' privacy.

One specific form of transparency that is currently widely adopted are transparency reports — regular voluntary public releases of information on content removals including the ones attributed to government requests (Hovyadinov, 2019; Parsons, 2019). Such reports serve the aforementioned purposes of transparency and are crucial for platform accountability because they inform the public about the platforms' content moderation practices and cooperation with various state actors. At the same time, such voluntary non-standardized reports allow platforms to engage in visibility moderation (Wagner et al., 2020) as by releasing certain information and framing it in a specific way in transparency reports, companies can promote a certain perspective on their platforms.

Depending on the specific information released in transparency reports, they can serve one or all of the aforementioned purposes of transparency – informing policy, empowering users, facilitating public opinion pressure or platforms' visibility management (MacCarthy, 2020; Wagner et al., 2020) – to varying degrees. Thus, we suggest the types of data disclosed by online platforms in their reports need to be scrutinized in order to establish whether and how they can meaningfully contribute to different transparency purposes. This is the issue that we address with the present paper. With it, we aim to make three main contributions. First, we conduct a comparative analysis of the data disclosed by major tech companies and social media platforms in their transparency reports with a particular focus on content moderation, relying on Santa Clara Principles (SCP) 2.0 — is a set of recommendations composed by scholars and human rights advocates to increase content moderation transparency. Second, we evaluate how helpful this data is in relation to the different purposes of transparency. Finally, we outline recommendations that could improve the level of transparency within the reports and beyond.

2. Related work

Transparency as an idea of making the world knowable can be traced back to the early Enlightenment (Gorwa & Garton Ash, 2020; Hood & Heald, 2006). However, it was the 20th century when the growing demand for accountability of state and corporate

entities resulted in the rapid increase in the number of legislative acts (e.g., Freedom of Information Act) as well as business practices (e.g., the adoption of environmental impact reports) which established transparency as an important component of societal regulation (Ananny & Crawford, 2018). It also led to the growing scholarly interest in the notion of transparency as well as its different forms (e.g., soft/hard or upwards/downwards transparency Fox, 2007; Heald, 2006).

The rise of online platforms prompted a new cycle of transparency-related debates. With tech companies promoting the ideas of openness and free access to information and platforms enabling new possibilities for whistle-blowing and initiating societal debates, transparency gained a new momentum in the platform-based media ecologies (Gorwa & Garton Ash, 2020). However, the functionality of the platforms themselves is not necessarily transparent, especially for the outsiders (Flyverbom, 2015) and in the cases of sensitive aspects of platform functionality such as content moderation (De Gregorio, 2020; Gorwa et al., 2020) or platform systems' functionality (Bastian et al., 2020; van Drunen et al., 2019). The discrepancies between the transparency ideal and the platform reality are further amplified by the evidence of increasing abuse of platforms' content moderation for censorship purposes by the authorities both in the democratic and the authoritarian contexts (Clark et al., 2017; Makhortykh, Urman, Wijermars, 2022).

These inconsistencies prompted the growing pressure on platforms to increase the transparency of their use of content moderation as well as their responsiveness to the authorities' requests. The first transparency reports from a digital platform was published by Google in 2010. In 2012, this example was followed by Twitter. In both cases, the companies focused on providing information on government requests related to content takedowns. The practice kept growing, in particular following the Snowden leaks (Parsons, 2019) which increased the public awareness about the potentially harmful effects of collaboration between the tech companies and governments.

The increasing use of transparency reporting has also prompted scholarly interest towards it. Parsons (2019) discussed the effectiveness of telecommunication companies' transparency reports in Canada to understand how they can promote changes in firm and government behavior. Heldt (2019) and Wagner et al. (2020) looked at transparency reporting in the context of the German Network Enforcement Act (NetzDG) and used its requirements as an analytical framework to discuss to what degree companies such as Facebook or Twitter meet the regulatory requirements. Kosta and Brewczyńska (2019) scrutinized the declared aim and the audience of examined transparency reports issued by 10 companies, including both major corporations (e.g., Facebook) and startups (e.g., Slack), whereas Hovyadinov (2019) examined transparency reports of Google, Facebook, and Twitter in the context of their cooperation with the Russian authorities' requests.

For the majority of the aforementioned studies, the assessments of transparency reports led to rather critical conclusions. One common source of criticism is the tendency of transparency reports to focus on aggregate level data (i.e., the overall number of content takedowns) without providing in-depth examination of specific cases, and the often obscure presentation of company policies leading to specific moderation decisions (Kosta & Brewczyńska, 2019; Suzor et al., 2019). Under these circumstances, despite the substantial volume of information provided in transparency reports, the public is not necessarily informed "about the real situation surrounding government access requests" (Kosta & Brewczyńska, 2019, p. 26).

Another common criticism of transparency reports relates to their tendency to focus on government requests at the expense of under-reporting companies' own actions in the context of content moderation (Hovyadinov, 2019; Kosta & Brewczyńska, 2019). The reasons for such an imbalance are not clear: while such reporting can inform the platforms' audience about the authorities' censorship efforts, it can also be used as a way of demonstrating platforms' compliance with the regimes (Hovyadinov, 2019) as well as "coaching governments on how to compel information from firms" (Parsons, 2019, p. 122).

This criticism of transparency reporting raises concerns not only about its limited effectiveness as an accountability mechanism, but also about the possibility of the reports serving as a form of "transparency-washing" (Zalnieriute, 2021) through the focus on procedural macro-issues to distract the public from the actual accountability-related matters. To address these shortcomings, a number of decisions were proposed ranging from better adapting transparency communication to specific audiences (Kosta & Brewczyńska, 2019) to facilitating access to data on content moderation (Suzor et al., 2019). In this study, we look at how one of the proposed solutions, namely the adoption of SCP 2.0, affects transparency reporting and their shortcomings.

3. Methodology

We have conducted a systematic comparative review of transparency reports released by 10 major technical companies and platforms owned by them (13 reports in total). In our analysis, we utilized the principles of a systematic literature review (Aromataris & Pearson, 2014), but instead of applying this technique for analyzing academic scholarship, we used it to examine transparency reports. First, we establish the criteria for the case selection (i.e., inclusion of specific transparency reports). Then, we describe analytical framework used to extract information from the reports according to set criteria (i.e., SCP 2.0 Principles, 2021). Finally, we synthesize all the findings in the form of a narrative summary followed by the discussion of their implications. The details on the case selection and analytical framework are presented below.

3.1. Case selection

We focused on the companies with the biggest reach in terms of the number of active users. Specifically, we included the following companies: Alphabet (Google), Apple, Microsoft, Amazon, Meta (formerly known as Facebook), Twitter, Bytedance (TikTok), Snap Inc (Snapchat), Pinterest, and Reddit. Our selection is comprised of the so-called Big Tech companies (PCMAG, 2021) and social media platforms with over 300 millions active users as of October 2021 (Hootsuite, 2021). We excluded the companies and platforms that provide no transparency reports. Among these were Quora as well as all non-US-based social media platforms except Bytedance:

WeChat and QQ (both owned by Tencent), Sina Weibo, Kuaishou and Telegram. Though we could not analyze transparency reports of these companies, we suggest that the very absence of such reports merits a separate discussion which we conduct in the respective section.

Several companies included in the analysis or their subsidiaries own and control multiple platforms.[1] It is the case of Google (YouTube, Google Search, Play Store), Apple (Apple, AppStore), Meta (Instagram, Facebook), Microsoft (Microsoft, Bing, LinkedIn, Github) and Amazon (Twitch). For Meta, Apple and Google there were single reports for all listed sites/platforms except a dedicated YouTube Community Guidelines enforcement report in the case of Google. Microsoft had a single report for Microsoft itself along with Bing, and separate reports for LinkedIn and Github. Amazon had separate reports for Amazon and Twitch. Additionally, in many cases individual reports consisted of several separate parts (e.g., Content Removal Requests; Digital Safety; Law Enforcement Requests Reports in the case of Microsoft). In such cases we treated all these multiple components as a single report as long as they pertained to the same (set of) platforms and reporting period. Hence, in total, we analyzed 13 distinct reports: Google and YouTube (Google, 2021), Microsoft (Microsoft, 2021a, 2021b, 2021c), Apple (Apple, 2021), Meta (Facebook, 2021), LinkedIn (LinkedIn, 2021), Github (Github, 2021), TikTok (TikTok, 2021), Snapchat (Inc, 2021), Pinterest (Pinterest, 2021), Reddit (Reddit, 2021), Twitter (Twitter, 2021), Amazon (Amazon, 2021) and Twitch (Twitch, 2021). In all cases we initially analyzed the latest reports available as of November 2021, complemented by an additional analysis of reports published between November 2021 and August 2022 (when available) that we conducted during the revision of this paper.

3.2. Analytical framework

For the analysis, we relied on the SCP 2.0 (Principles, 2021) as our main framework. The SCP originally were formulated and published in 2018 by a group of academic experts, human rights organizations and advocates with the aim of enabling possibilities to "obtain meaningful transparency and accountability around Internet platforms' increasingly aggressive moderation of user-generated content" (Principles, 2021). Since then, several major companies including some of those whose reports are included in our analysis have endorsed the SCP.

In 2021, an updated version of the SCP was published after consultation with experts across the globe that highlighted the inequities in transparency reporting in different national contexts (Principles, 2021). As SCP creators note, they aimed to create an aspirational standard in terms of the information that the companies need to disclose in order to enable meaningful transparency and accountability (Principles, 2021). For exactly this purpose, there is a *Numbers* Section in SCP that outlines concrete recommendations about the data that needs to be included in Internet companies' transparency reports. The parameters outlined in this section that form the aspirational standard of transparency reporting in the context of content moderation are at the core of our analysis. Specifically, we analyzed the latest available (as described above) transparency report from each company or platform (e.g., if a company owns several sites and has separate reports or parts of reports for each of them such as in the case of separate reports for Facebook and Instagram owned by Meta) and noted whether the given report contains the information on each of the parameters outlined in the *Numbers* section of the SCP. This section of SCP has three subsections: content and accounts actioned *without* the involvement of state actors; decisions made *with* the involvement of state actors; flagging processes. In SCP 2.0 the involvement of state actors is defined as one of the following: "state's involvement in the development and enforcement of the company's rules and policies, either to comply with local law or serve other state interests" or direct demands/requests from state actors — e.g., to remove certain content (Principles, 2021). Importantly, as in certain cases local legislation might contradict company's internal rules, SCP highlights that users should be able to access "details of the process by which content or accounts flagged by state actors are assessed, whether on the basis of the company's rules or policies or local laws" (Principles, 2021).

We organized our analysis and results according to the three subsections of SCP, splitting the results section into three corresponding sections. In addition to evaluating the reports for the presence of data included in the Numbers section of SCP, we noted all the additional information included in each report (i.e., the parameters not mentioned by the SCP but covered by the report). Afterwards, we synthesized all the findings and examined the trends that emerge from the analysis in terms of the similarities and differences across the reports and their correspondence to the SCP. We then contextualized the findings connecting them to the purposes of transparency described in the Introduction — informing policy, empowering users, facilitating public opinion pressure or platforms' visibility management (MacCarthy, 2020; Wagner et al., 2020).

4. Results

The structure of this section follows the structure of the *Numbers* section of the SCP. There are five subsections, one for each of the three sets of parameters that the SCP postulate – content and accounts actioned without the involvement of state actors; decisions made with the involvement of state actors; flagging processes, – followed by a subsection on the additional data that can be found in the companies' reports but is not mentioned in the SCP, and a summary in the end.

[1] We follow the definition of platforms given by Gillespie (2018): online sites and services that: (a) host, organize, and circulate users' shared content or social interactions for them, (b) without having produced or commissioned (the bulk of) that content, (c) built on an infrastructure, beneath that circulation of information, for processing data for customer service, advertising, and profit.

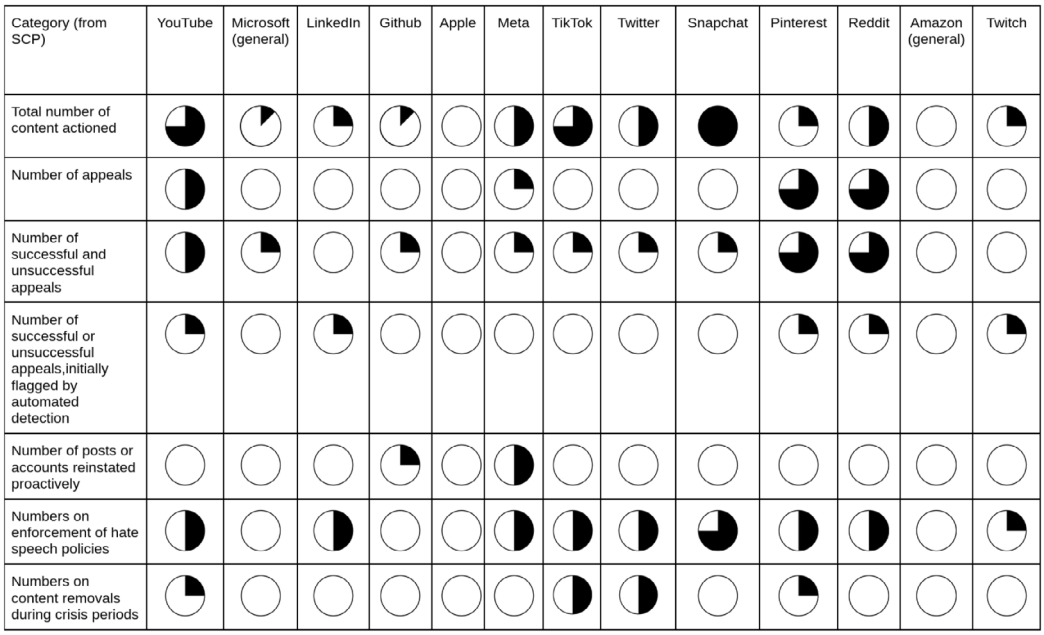


Fig. 1. Visual overview of the company reports' levels of compliance with SCP on content sanctioned without the involvement of state actors. Full (black) circles correspond to full compliance with SCP, empty (white) circles mean that the data on a given type of information is not provided in a report at all.

4.1. Content and accounts actioned without the involvement of state actors

The SCP suggest that companies should release the data on the pieces of content and user accounts that were subject to different actions from the platform – e.g., those removed or reinstated, – without any involvement from state actors. The SCP advise that each category of data published should be broken down by country or region and whenever possible by the type of violation. We go through the data types listed in the SCP one by one and summarize whether and how corresponding numbers appear in the companies' transparency reports. For the convenience of the reader, we also provide a visual overview of the degree to which individual companies' reports comply with the standards set in SCP for each parameter related to the content/accounts actioned without the involvement of state actors in Fig. 1.

4.1.1. Total number of pieces of content actioned and accounts suspended

The only report that contains this information in full accordance with what is suggested by the SCP is that of Snapchat. The reports of almost all other companies comply with the suggestions but only partially.

Most reports mention the total number of accounts and posts removed or suspended with a breakdown by the policy violation but lack detailed breakdown by the country/region. TikTok and YouTube[2] provide country-level breakdown only for pieces of content, not for accounts. Twitter, Reddit and Meta provide country-level breakdown only for government-related requests or legal violations (see the next subsection) but not for the actions without the involvement of government actors. Pinterest, Twitch and LinkedIn list only the total aggregate numbers without any country-level breakdown.

The transparency report by Microsoft is rather rigid. It includes the data taken down only in relation to specific issues, such as the non-consensual intimate imagery, terrorist content, child sexual exploitation and abuse imagery. There is no breakdown by the country for the content removed in relation to these violations. It is also unclear if any content was removed by Microsoft in relation to other issues and policies. Similar applies to Github that lists the total number of content restricted for the violation of its Community Guidelines. However, the report does not include numbers corresponding to the spam and malware-related violations. Hence, the numbers provided relate only to a limited set of Github's policy violations and the total amount of content removed is unknown. Finally, Apple and Amazon do not provide data on the content taken down by the companies themselves (i.e., without requests by governments or law enforcement agencies).

[2] In this subsection we discuss only YouTube's Community Guidelines enforcement report but not other reports by Google because YouTube's dedicated report is the only one that contains any information on content removals without the involvement of state actors. Google's other reports provide information on state actors' involvement and are discussed in the next subsection.

4.1.2. Number of appeals of decisions to action content or suspend accounts

No transparency report contains information on this in full accordance with the SCP.

Microsoft, Github, LinkedIn, Twitter, Snapchat, Amazon, Twitch and TikTok list no information on the total number of appeals. Meta lists the total number of appeals only for some policy violations (e.g., hate speech) but not others, and does not include breakdown by country. Pinterest, Reddit and YouTube's reports contain information on the number of appeals but without breakdown by the country; further, in the case of YouTube this data is available only for individual pieces of content (videos) but not the accounts. Apple again does not provide this information for the content taken down without the involvement of state actors and it is only available for the content removed upon legal requests.

4.1.3. Number (or percentage) of successful appeals that resulted in pieces of content or accounts being reinstated, and the number (or percentage) of unsuccessful appeals

No report contains information on this in full compliance with the SCP.

Similarly to the data on the total number of appeals, Pinterest, Reddit and YouTube reports contain information on the number of successful/unsuccessful appeals but without breakdown by the country. In the case of YouTube, these data are available only for individual pieces of content but not accounts, and without the country-level breakdown. TikTok includes the data on the number of successful appeals only, broken down by country, but not on the number of unsuccessful ones. Twitter, Github, Snapchat and Microsoft list only shares or numbers of content that was reinstated out of the content removed for violating specific policies; it is unclear, however, whether these reinstatements were done proactively or after appeals, and no data on appeals is included. Meta provides only the number of successful appeals for some policy violations but not the others and does not include breakdown by the country. Apple again includes this data only for the legal requests-based removals but not for those which do not involve state actors. Other reports do not include information related to the successful/unsuccessful appeals at all.

4.1.4. Number (or percentage) of successful or unsuccessful appeals of content initially flagged by automated detection

Reports from LinkedIn, YouTube, Reddit, Twitch and Pinterest provide the numbers of content initially flagged by automated detection that was removed, but no information on related appeals is provided. Other companies' reports do not include any data on this.

4.1.5. Number of posts or accounts reinstated by the company proactively, without any appeal, after recognition that they had been erroneously actioned or suspended

The only company whose report partially addresses this point is Meta. This information is included in relation to some policy violations but not the others, and without breakdown by country. Github's report includes data on the total number of reinstatements, without further breakdown. Other platforms' reports contain no data on this.

4.1.6. Numbers reflecting enforcement of hate speech policies, by targeted group or characteristic, where apparent

None of the companies include breakdowns by the targeted groups or characteristics. Snapchat is the only one to provide information on hate speech violations-related rule enforcement broken down by country. Meta, LinkedIn, Twitter, YouTube, TikTok, Reddit and Pinterest reports contain information on the total numbers of posts removed for violating hate speech-related policies but without breakdowns the by targeted groups/characteristics and by the country/region. Twitch includes information on the aggregate number of removals related to hate speech violations without any breakdown by country and only for content that was reported by users. Microsoft, Apple, Github and Amazon provide no data on this.

4.1.7. Numbers related to content removals and restrictions made during crisis periods, such as during the COVID-19 pandemic and periods of violent conflict

Most companies do not provide such crises-specific data on the content removals and/or restrictions. The exceptions are the reports from YouTube, Twitter, TikTok and Pinterest that include dedicated sections about COVID-19-related content removals and/or extended explanations on the way a given company tries to counter COVID-19-related misinformation. Additionally, Twitter and TikTok include dedicated sections on the elections-related content moderation.

4.1.8. Summary

In principle, data on the content and accounts actioned by a platform without any involvement of the state actors should shed light on the moderation processes that are solely based on the platform policies. This is especially crucial for platform accountability in the countries with "broad immunity" and "conditional liability" approaches to platform liability such as the US. In practice, no company provides information in full compliance with the SCP. There is a lot of divergence across the reports in the level of compliance with these suggestions. It is unclear how the companies decide on which information to provide in their reports on content moderation without the involvement of state actors, however the absence of standardized requirements in this case is definitely conducive to the utilization of transparency by them as visibility management and engagement in "transparency washing" (Zalnieriute, 2021).

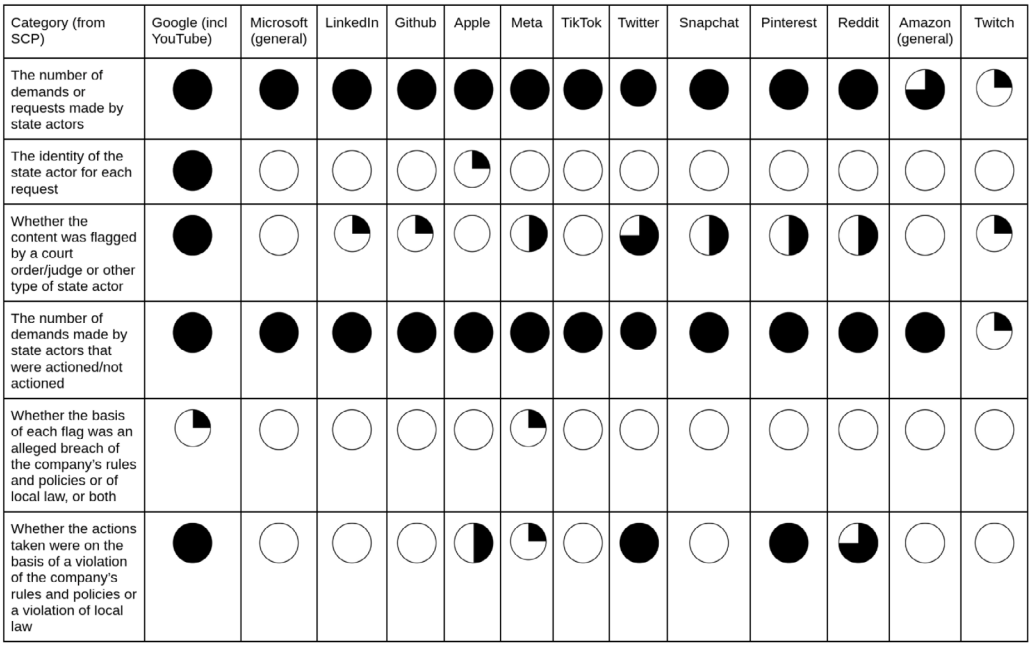


Fig. 2. Visual overview of the company reports' levels of compliance with SCP on content sanctioned with the involvement of state actors. Full (black) circles correspond to full compliance with SCP, empty (white) circles mean that the data on a given type of information is not provided in a report at all.

4.2. Decisions made with the involvement of state actors

The SCP outline "special reporting requirements" when it comes to the decisions made with the involvement of state actors. As with the suggestions regarding content moderation without state actors' involvement, relevant data on such decisions should be broken down by country. We present a visual overview of the degree to which individual companies' reports comply with the standards set in SCP for each parameter related to the decisions made with the involvement of state actors in Fig. 2.

4.2.1. The number of demands or requests made by state actors for content or accounts to be actioned

All transparency reports[3] analyzed with the exception of Twitch contained this information broken down by the country. In the case of Amazon this applies only for requests for the user data but not for the content removals. The report by Twitch provided only information on subpoenas and preservation holds that was processed by the company. This data had no breakdown by country, and it is unclear if the data presented corresponds to all types of government requests — i.e., whether there were requests other than subpoenas and preservation holds.

4.2.2. The identity of the state actor for each request

Only Google transparency report had this data provided. Apple report had this information only for the US National Security-related requests. None of the other reports contained any data on this.

4.2.3. Whether the content was flagged by a court order/judge or other type of state actor

Google's report contains this information in full compliance with the SCP, broken down by country. Meta, Snapchat and Reddit provide such data only for the US but not other countries. LinkedIn and Github list this information only for the US-based requests for user data. Twitter provides fine-grained data for the US with breakdown by the multiple categories of legal requests such as court orders, subpoenas, search warrants and other. For other countries, the data is broken down only between the two categories — court order vs other requests. On Pinterest the distinction similar to the one on Twitter is available for the government requests for providing account information but not for the content removals. Twitch includes information on subpoenas and preservation holds processed, however it is not broken down by country, and there is no information about other types of requests. Other reports do not include information on this.

[3] In this section, when discussing Google, we refer only to the platform's overall reports that in the case of decisions made with the involvement of state actors apply to the whole range of the company's products including Google Search and YouTube. It is unclear whether the report also covers PlayStore.

4.2.4. The number of demands or requests made by state actors that were actioned and the number of demands or requests that did not result in actioning

All transparency reports contain this information, with the exception of Twitch. Twitch lists only the number of subpoenas and preservation hold requests the company "processed". It is unclear, whether "processed" means "actioned" and if so, whether there were additional requests not actioned by the company.

4.2.5. Whether the basis of each flag was an alleged breach of the company's rules and policies (and, if so, which rules or policies) or of local law (and, if so, which provisions of local law), or both

None of the reports contain such a breakdown by local laws vs company policy violations when it comes to the requests involving state actors. However, whenever data on removal requests involving state actors is provided, these are typically listed as "legal requests" which might imply that all such requests are made on the basis of local laws. Thus, the reports' treatment of this component is rather ambiguous.

When it comes to additional details, Meta provides information on the number of content requested to be removed for violation of specific laws in each country. Google lists the number of requests per country that were filed for specific reasons such as "government criticism", "copyright", "impersonation" and other broadly defined reasons that can correspond to specific local laws. Other companies' reports do not provide such information.

4.2.6. Whether the actions taken against content were on the basis of a violation of the company's rules and policies or a violation of local law

Google, Twitter and Pinterest provide this information with country-level breakdown. Reddit's report includes this data on aggregate level only, without breakdown by country. Apple provides data on the app removal requests granted on the basis of platform policies vs legal violations. Meta's report does not explicitly address this point but as noted in the previous subsection the company lists the number of the cases where content was removed or withheld in a country for violating local laws. Whether such content also violated company's policies however is not explicated for this data. Other reports do not address this.

4.2.7. Summary

Overall, when it comes to the decisions made with the involvement of state actors, the examined transparency reports fulfill the SCP suggestions better than in the case with the companies' internal moderation practices. Most reports contain information suggested by the SCP 2.0 at least on aggregate level. The coverage on additional details such as the basis of requests/content removal or the identity of state actors filing requests varies across reports. All of them, nonetheless, are somewhat conducive to meaningful transparency in the context of empowering users to make informed decisions and enabling public pressure if not on platforms themselves, then on the governments, thus contributing to the transparency on state actions.

Google's reports come closest to fulfilling the SCP in this section. However, the fact that Google does not provide a breakdown across its wide array of products (e.g., Search, PlayStore, Google Docs, Google Drive, etc.), introduces additional opacity.

4.3. Flagging processes

Flagging is a mechanism that allows users to directly report offensive content to online platforms and their moderators. While the idea behind flagging is to empower users to combat offensive content and harassment including that targeted at them personally, flagging processes can also be abused and utilized to harass and silence users that did not violate any platform policies or norms (Crawford & Gillespie, 2016). Citing concerns regarding the potential misuse of flagging processes against other users, the SCP suggest that the companies should publish numbers regarding flagging as to shed light on the potential scope of abuse of this mechanism. The SCP suggest this data should be disaggregated by country or region. In Fig. 3 we present a visual overview of the degree to which individual companies' reports comply with the standards set in SCP for each parameter related to the flagging processes.

4.3.1. The total number of flags received over a given period of time

Snapchat is the only one to provide this data broken down by country, however it is ambiguous whether this data encompasses all types of flags or only those received from users. Twitter, Reddit, Twitch and Pinterest include the total numbers of reported content but do not disaggregate it across countries or regions. In the cases of Twitter, Pinterest and Twitch, similarly to Snapchat, it is unclear whether the data includes all flags or only those from the users.

Meta, TikTok, Github, LinkedIn, YouTube[4] and Amazon do not provide information on the total number of flags. Apple lists the number of "Private Party" (content removal) requests only in the US, it is however unclear if this refers to flagging or to legal requests from private parties. Microsoft includes only the number of flags for non-consensual sexual imagery policy violation, without breakdown by country.

[4] In this subsection we again refer only to YouTube Community Guidelines enforcement report as more general reports by Google for other products do not have relevant information.

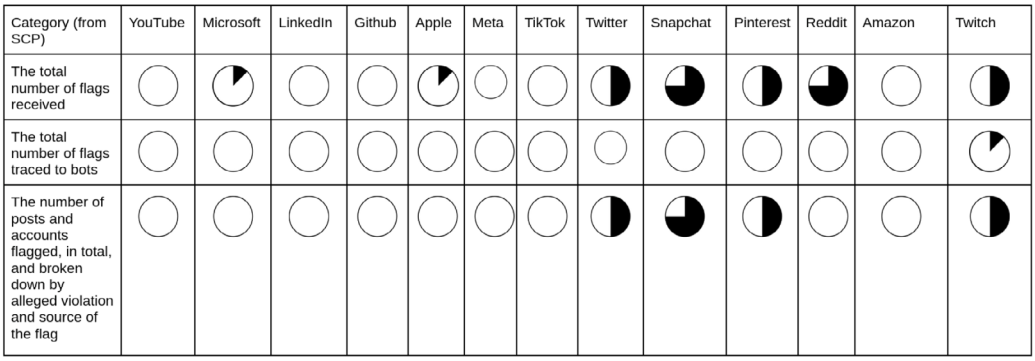


Fig. 3. Visual overview of the company reports' levels of compliance with SCP on flagging processes. Full (black) circles correspond to full compliance with SCP, empty (white) circles mean that the data on a given type of information is not provided in a report at all.

4.3.2. The total number of flags traced to bots

None of the reports includes this information. In the case of Twitch, the fact that some flags were traced to "brigades"[5] is mentioned in the text of the report, but no related numbers are provided.

4.3.3. The number of posts and accounts flagged, in total, and broken down by alleged violation of rules and policies and source of the flag

Snapchat's report lists the number of reports within the in-app reporting system broken down by policy violation and country. Twitter, Twitch and Pinterest include the data on the number of user reports per policy violation but not by country. Other companies' reports do not mention this information.

4.3.4. Summary

Generally, the reports' coverage of numbers related to flagging processes is the least compliant with SCP suggestions. Snapchat's report comes the closest to addressing all points formulated in the Principles, followed by Twitter, Twitch and Pinterest. The opacity with regard to the flagging processes remains high across platforms, especially when it comes to the potential of (automated) abuse of flagging. In the current state, the transparency reports fall short of facilitating meaningful transparency on flagging processes for empowering users or informing policy decisions.

4.4. Additional information included in company reports beyond what is suggested by the Santa Clara Principles 2.0

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. Naturally, regardless of the level of compliance with the SCP, specific reports can contain information that goes beyond what is stipulated by the principles. All reports that we examined included some data that goes beyond the suggestions of the SCP. Here we briefly outline what types of such data can be found in different reports.

All examined reports with the exception of Twitch provided **data on government requests for user information and on the level of company compliance with such requests**, broken down by country. Depending on specific reports, some additional related data was available. For instance, Amazon explicated the share of "content" and "non-content" information in the data the company disclosed to the government (i.e., "non-content" information relates to basic account information, while "content" information can include users' communications and other detailed data), while Github's report includes the number of cases when users were or were not notified about the legal request data disclosures. Most companies also explicitly mentioned the number of user data requests that came from the US National Security Agency. Twitch's report was the only one to include just "subpoenas and preservation holds processed", as discussed in the previous sections, but not other types of government requests. However, it was also the only report to mention the escalations to law enforcement triggered by the platform itself after identifying potentially illegal content. Other platforms' reports contained no information on this, though it is unclear whether they never escalate to law enforcement or just do not provide the data on this.

Another type of data commonly found across the reports is information on content takedowns and sometimes on the takedown requests in relation to **intellectual property protection**. This was included in all reports with the exception of those by Pinterest, Twitch and Amazon.

Some reports include companies' **insights on certain events or practices that occur beyond the companies' platforms**. For example, Meta and Google provide information on the internet disruptions across the world that were detected by these companies, while Twitter includes data on different email providers' (those not affiliated with Twitter) security and privacy practices.

[5] In this case, groups of users who collectively, in a coordinated manner, report (flag) another user(s)' accounts or posts for alleged rule violation – typically, in absence of an actual violation, – to get the platform to suspend the target user's account.

When it comes to other types of data, there are little similarities across the reports. Reddit and TikTok mention the number of cases when content was removed as a result of automated flagging. In fact, Reddit lists data on content removed by different types of actors (automated moderation, human moderators, admins); as well as on content removed according to each Reddit content policy violation type and type of removed content (e.g., post/comment/subreddit) and types of actions taken (e.g., temporary or permanent bans) against specific accounts and/or posts broken down by Reddit content policy violation. TikTok discloses the number of ads rejected for violating company policies. Twitter gives data on the malicious automation and state-backed information operations detected on the platform, as well as provides statistics on how users protect their accounts. LinkedIn lists data on the number of fake accounts and spam detected. Meta's report includes data on content that was viewed by a high number of Facebook users, while that by Pinterest provides estimated number of people who saw content that was later taken down for violating company policies', broken down by policy. Twitch provides details on the ratio between different content moderation-related enforcement actions and number of hours of video watched on the platform. Additionally, information on the share of chat messages removed automatically vs through human content moderation is listed. Finally, Google's reports include detailed information on the content removed under German Network Enforcement Law and, separately, additional details on specific government removal requests that Google deems to be of "public interest", broken down by country. Though it is unclear how the company determines whether something is of public interest or not, this, along with the initiatives such as Lumen Database archiving of government requests in which Google and Twitter participate, can help shed further light on the nature of governments' requests and companies decision-making in response to that.

4.4.1. Report design

There are currently no suggestions regarding the design of reports outlined in the SCP. Nonetheless, previous research shows that design choices can have major implications for the usability of specific reports and thus reduce or increase their overall opacity (Wagner et al., 2020). While we did not examine such design choices in great detail as that would merit a separate UX-focused analysis that is out of scope of our paper, we have recorded and analyzed more general characteristics of the examined reports such as the formats in which the reports and corresponding data is available.

The majority of reports are presented in the form of web-based textual summaries with accompanying graphs and tables. This is the case with all reports except ones Google/YouTube, Microsoft, Meta, Apple and Twitter that also have interactive elements in the form of web-based dashboards. What perhaps is more important in the context of transparency for external users and potential auditors is the availability of machine-readable files that could be used for subsequent analysis of the data from companies' reports. Most companies provide such files (in .csv or .xlsx formats) though not necessarily for all the types of data included in their transparency reports. But there are cases where the data in machine-readable format is not available at all. Among the examined reports this applies to Github, LinkedIn, TikTok, Pinterest, Twitch and Amazon. Such non-inclusion of machine-readable data decreases the usability of the reports for external analysis.

Finally, given the global reach of tech companies, it is important to examine whether their transparency report data is accessible to users in non-English-speaking countries they operate in. We have also recorded and summarized information on this. Only some companies among those examined had their reports translated into languages other than English: Google, TikTok, Pinterest, Twitter, Reddit, Twitch, Snapchat, with Google and Snapchat having translations in the widest array of languages. The report by Twitch had the main text of the report translated into multiple languages but in all the translations the text in the Figures included in the report remained in English. The reports by "Big Tech" companies except Google (those of Amazon (main), Apple, Microsoft (incl. Github, LinkedIn), Meta) were published in English only, though the reach of these companies and their products undoubtedly goes beyond English-speaking audiences.

4.5. Summary

We observe great variance in the level of compliance to SCP suggestions across different reports and reporting dimensions. The compliance with reporting criteria related to decisions involving state actors tends to be the highest across companies, while internal moderation decisions and flagging processes on the platforms remain more opaque. Among the data not included in SCP but still reported by the companies, almost all companies include additional data on practices related to intellectual property protection and the US National Security-related requests. We discuss possible reasons and implications of this and other observations in the next section.

5. Discussion

Our analysis has demonstrated that there are vast discrepancies across companies in the data they provide in their transparency reports, with none of the examined reports being fully compliant with the SCP 2.0. For a brief overview of the level of compliance – and corresponding divergence across the reports – one can refer to the visual summaries in Figs. 1, 2, 3. While several studies have empirically examined the content of transparency reports before, they looked at fewer platforms and/or had a more narrow focus than our analysis — e.g., Kosta and Brewczyńska (2019) focused on reporting about government requests, whereas Hovyadinov (2019) examined reporting in the context of Russia. To the best of our knowledge, our paper is the first one to include an empirical analysis that includes a wide range of platforms and focuses not just on one aspect of reporting – e.g., government requests – but also discusses companies' own moderation practices and flagging processes.

We found that companies tend to report more data and in greater detail on the moderation decisions involving state actors than on their internal moderation practices, in line with previous observations about transparency reports (Hovyadinov, 2019; Kosta & Brewczyńska, 2019). This can be attributed to the fact that many companies started publishing the voluntarily produced reports after – and possibly as a reaction to – the Snowden leaks that increased the public awareness about state surveillance (Gorwa & Garton Ash, 2020). Seemingly, companies are more willing to increase transparency on their relations with the various states and their demands rather than on their own moderation practices, with the latter remaining largely obscure. We argue that this tendency is problematic because currently transparency reports serve more to the companies' visibility management (Wagner et al., 2020) than to the purposes of meaningful transparency that we discussed in the Introduction (i.e., informing policy, empowering users or facilitating public opinion pressure MacCarthy, 2020).

One solution to that could be the introduction of a dedicated legislation that would require companies to disclose certain information for transparency purposes. Such requirements were introduced in the Digital Services Act (DSA) - the regulation of online intermediaries within the EU (European Commission, 2020) that received final approval from the European Council on October 4, 2022. In the context of transparency reporting, DSA contains multiple provisions that can be seen as the first step to standardized transnational regulation of transparency reporting. Further, it is likely that DSA will influence transparency reporting legislation in countries beyond the EU, just like GDPR influenced data privacy laws across the globe (Greenleaf, 2022). For this reason, when discussing our recommendations regarding transparency reporting based on our analysis in the next few sections, we will also contextualize them with regard to the DSA. Importantly, similarly to the DSA that does not put any reporting requirements on smaller companies, our suggestions also relate primarily to established big platforms. Some of the suggestions might be not feasible or counterproductive for smaller companies and might obstruct their entrance to the market and growth, — so we underscore that what we outline below concerns only major international platforms.

Though we discuss the points below primarily in the context of legislation, some of them – i.e., those that can be implemented by a single company without coordinating with all the others – can also be regarded as recommendations to individual companies as a way of improving their current transparency reporting practices. Just like releasing transparency reports per se can help companies build trust with their user base, making the reports more accessible and conducive to meaningful transparency can foster such trust as well. Voluntary reporting by the companies is all the more important given that legislation that would mandate such reporting and set standards is related to inherently complex issues. For example, it is unclear how platforms would go about contradicting reporting requirements in the legislations of different countries, should such a case arise.

5.1. Towards meaningful regulation of transparency reporting

5.1.1. Release of transparency reports

Transparency reporting-regulating legislation, though not without limits depending on the implementation as the example of German NetzDG shows (Heldt, 2019; Wagner et al., 2020), could first of all enforce the release of transparency reports. As noted in the Methods section, at this point not all companies, especially those with headquarters outside the US, even produce such reports. It is unclear why non-reporting is more frequent in the case of non-US-based companies. One reason could be the decreased societal/institutional pressure on companies outside the US. Regardless of the reasons, however, the introduction of legislation on transparency reporting could force these companies to adopt transparency reports. The legislation could also outline the frequency with which the reports are released to increase their comparability across companies, because currently some companies produce reports once and some twice a year.

The DSA proposal currently includes such requirements, so when it goes into force, the large online intermediaries will be forced to regularly – at least once a year – release transparency reports. We suggest this is a meaningful step towards greater transparency, from which will benefit not only users in the EU but also in other countries. Even if companies end up releasing the data only related to their activities within the EU, the need to do it for the large platforms that currently do not publish any transparency reports will definitely help decrease the opacity of their operations.

5.1.2. Report design and access to machine-readable data

Previous analysis of reporting practices based on NetzDG has revealed that the designs of transparency reports vary across the companies (Wagner et al., 2020). We have also observed discrepancies in the report design even on a macro level. Importantly, not all companies provide access to machine-readable data from the reports they release. We suggest that the legislation could explicitly outline the necessity to release such data to facilitate its analysis by external auditors, users or researchers. The final version of the DSA, for instance, includes a requirement that reports should be published in a machine-readable format (European Parliament, 2022). It is, however, not clarified currently what the machine-readable format should be. Nonetheless, according to the Articles 13 and 23 of the DSA, the Commission will be able to "adopt implementing acts to lay down templates concerning the form, content and other details of reports" (European Parliament, 2022) — in other words, specify details such as the specifics of the machine-readable format through implementing acts at a later point. If all companies release data in vastly different formats, it will be difficult for external auditors and researchers to conduct comparative analysis on it. Hence, we suggest it would be beneficial if relevant legislation requires that all companies release standardized data making their reports more comparable and thus more usable, because currently the types and formats of data released by the companies are inconsistent and thus difficult to compare (Wagner et al., 2020). Such standardized reporting in machine-readable format is key to the purpose of transparency that deals with informing policy decisions as it will facilitate access to the data necessary to conduct comparative analyses and make conclusions about the current state of content moderation and problematic content across platforms. The results of such analyses

can then enable better-informed policies and regulations with regard to these issues. Besides, the standardized requirements for information reported will decrease the ability of platforms to utilize transparency reporting for visibility management rather than providing meaningful transparency. Finally, we suggest that in the specific case of the DSA, the potential definition of machine-readable format requirements through additional implementing acts and not in the original DSA text will allow for greater flexibility and will make it possible to regularly amend the requirements if necessary so that the outlined format does not become outdated.

5.1.3. Context-aware international reporting and additional details on socially and politically important cases

We suggest that the potential legislative regulation of transparency reporting, even if drafted and enforced in the best way possible, would hardly be enough to increase overall transparency of tech companies if done on a national level only. This again is demonstrated with NetzDG, where the companies complying with it tend to release the required data only for Germany but not other countries (see, for instance, the dedicated NetzDG section of Google report). Hence, DSA is a welcome addition as it will govern reporting across countries and might influence relevant legislation across the globe. That being said, transnational regulation is a complex issue in itself, and merits a separate discussion on how it could be implemented. Specifically, even consistent reporting of data across countries might not be meaningful given vast political and socio-economic differences. For instance, reporting on content moderation decisions with the involvement of state actors would likely need to be done differently in the case of democratic countries with independent judicial systems and non-democratic countries where judicial systems are dependent on the executive powers.

In the latter case transparency about certain decisions involving the state actors can be especially important in the context of transparency's purpose of empowering users and influencing public opinion since compliance with some of their demands might result in companies' infringing on their users' rights such as the right to free speech. For instance, human rights groups have criticized Apple's and Google's removal of a voting advice app in Russia amid 2021 elections and Apple's removal of an app used by protesters in Hong Kong in 2019 (Post, 2021). Though in both cases the removals were allegedly done under pressure from the autocratic regimes and contributed to the regimes' efforts in silencing dissent, there is little clarity on how exactly the decisions were made. In such cases simple reporting of statistics related to the authorities' requests is arguably insufficient, because it obscures the basis of content removal and the decision-making process related to that, which can be crucial information if the users' rights are infringed. The importance of transparency is exemplified by the case of the Russian voting advice app as it was reinstated by Google after the elections, making the reasons for both removal and reinstatement critical for the external assessment of how legitimate and well-grounded company's decision to take it down was in the first place.

The differences in transparency metrics that can be meaningful in various political regimes merit a separate in-depth discussion that is out of scope of the current paper. However, we find it important to highlight that transparency reporting requirements might need to be different depending on the context. In the case with authoritarian regimes, for instance, reporting on state requests and cooperation with the governments should be more in-depth than simple statistics and should provide detailed explanations about the way companies made decisions regarding such cooperation. Google has made a step in this direction by releasing some details on state actors' requests that the company deems "of public interest"[6] (Google, 2021). It is nonetheless not clear how the company decides what is of public interest and what is not, and a publication of guidelines that Google uses to determine this would increase the reliability of such details.

Another step which is useful for increasing meaningful transparency and is taken by both Google and Twitter is to report the details on state requests to a dedicated independent database – Lumen Database (Lumen, 2021) – that is publicly available and hosted by Berkman Klein Center. We suggest that such direct reporting of requests increases transparency in the companies' dealings with state actors to a greater extent than providing statistics alone, and argue that potential legislation governing companies' transparency should consider including such transparency-increasing measures. The DSA contains a similar – and even broader – requirement for the platforms to submit the "decisions and statements of reasons of the providers of online platforms when they remove or otherwise restrict availability of and access to content" (European Parliament, 2022) to a special database that will be published and maintained by the European Commission. Such a database will be conducive to meaningful transparency and will help inform policy and academic debates around content moderation as well as provide users with additional contextual information that can help them to make informed decisions about the usage of specific platforms.

5.1.4. Accessibility to international audiences

Given the global reach and scale of technical companies' operations, transparency reporting should be done on international level with comparable data being available across countries. This data should also be available in the national languages of all countries in which a company operates so that local users can easily access information about their country. Currently, this is not the case. We suggest that translating existing reports into other languages is not too difficult or costly and would increase their accessibility to the users across the globe. Hence, the presence of such translations could be a reasonable requirement for all transparency reports. The DSA requirements in the current form do not make it obligatory to publish transparency reports in all languages of the EU: unlike with the Terms and Conditions that, according to DSA, need to be translated into all EU languages, the Reports can be published in one of the languages of the Union only (European Parliament, 2022). We suggest that in the context of regulation within Europe this is an insufficient measure. While publishing the reports in only one EU language might help inform policy decisions, it limits the users' ability to get information about the platforms' decisions and thus hinders those purposes of transparency that deal with the users rather than regulators or experts.

[6] Google provided vaguely worded explanation on the aforementioned case citing the grounds that Russian authorities used to request removal of content and mentioning that the content was blocked locally and reinstated after the election period. The decision-making process, however, was not explicated (Google, 2021).

5.1.5. Reporting requirements for companies that own multiple platforms

Generally, we observe that the "Big Tech" companies tend to release more opaque transparency reports than the smaller companies — at least among the reports examined. For example, Apple does not release information on content removals without the involvement of state actors, thus company's internal moderation practices – e.g. those applied to its App Store – remain opaque. Amazon does not provide any information on content removals at all, including those done according to the state requests. And while on Amazon's platforms, such as Amazon's marketplace, moderation arguably plays less of a role than on the other types of platforms, media reports suggest that moderation upon state requests occurs too: for instance, Amazon reportedly removed negative comments about Xi Jinping's book upon a demand from Chinese government (Stecklow & Dastin, 2021). Such decisions are currently opaque, because Amazon provides no related data. Similarly, a high degree of opacity – whether intentional or not – is characteristic of Google's reports. Though the company was a pioneer in transparency reporting and with respect to the decisions involving state actors its reports come closest to fulfilling the SCP, the high degree of opacity comes from the fact that the state-related reports are not broken down across Google's products, and the reports on internal moderation practices and flagging processes are available only for YouTube. Thus, it remains unclear how Google moderates platforms such as Google Search, Play Store or Google Docs. Similar absence of breakdown by the product is characteristic of Microsoft reports. Based on these observations, we suggest that in the case of big companies that own a wide array of products, transparency reports should be issued with a clear breakdown across the products to increase their usability.

5.1.6. Summary

Our suggestions with regard to the reporting standards necessary to implement for more meaningful transparency can be summarized along three dimensions: accessibility, harmonization, more detailed reporting. Below, we concisely list all suggestions for each of these dimensions that we discussed in detail above.

- **Accessibility:** release of reports; release of relevant data in machine-readable format; translation of reports into the national languages of all countries in which a company/platform operates.
- **Harmonization:** standardized report release frequency across all platforms/companies; standardized machine-readable data format for relevant data releases; same transparency reporting standards for all platforms owned by each company, with a clear platform-level breakdown within each report.
- **More detailed reporting:** detailed explanations of decision-making processes behind the decisions made with the involvement of state actors; submission of additional information on the decisions made with the involvement of state actors/on state requests to an independent external database (such as Lumen).

6. Conclusion

In this paper we have conducted a comparative analysis of transparency reports by major tech companies using Santa Clara Principles 2.0 as the main analytical framework. Our analysis has revealed that none of the reports fully adhere to the suggestions of the SCP, with the gaps being particularly pronounced in the case of reporting on companies' internal moderation practices and flagging processes rather than decisions involving state actors. We have also observed that in many cases the so-called "Big Tech" companies' reports tend to be more opaque and rigid than those of the smaller companies. We have summarized our observations and then discussed their implications for potential legislative regulation of transparency reporting, contextualizing those against the Digital Services Act within the EU. One major limitation of our study is that it is based only on the most recent reports from each company, and thus relies on a single observation point. However, transparency reporting practices constantly evolve, and tracing their evolution across companies would be a worthwhile task for the future research. This being said, such an analysis is not always possible — in the case of companies that publish their reports in the form of web-based dashboards (e.g., Meta), it is unclear at what time point certain elements of said dashboard became a part of the report, thus any future research aiming to trace the evolution of such reports comprehensively will either need to rely on web archives when available or conduct interviews with company representatives.

Data availability

We build on openly accessible data sources and reference all of them in the manuscript.

References

Albu, O. B., & Flyverbom, M. (2019). Organizational transparency: Conceptualizations, conditions, and consequences. *Business & Society*, *58*(2), 268–297.

Amazon (2021). Law Enforcement Information Requests - Amazon Customer Service. URL: https://www.amazon.com/gp/help/customer/display.html?nodeId=GYSDRGWQ2C2CRYEF.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989.

Apple (2021). Privacy - Transparency Report - Apple. URL: https://www.apple.com/legal/transparency/.

Aromataris, E., & Pearson, A. (2014). The Systematic Review: An Overview. *AJN the American Journal of Nursing*, *114*(3), 53–58. http://dx.doi.org/10.1097/01.NAJ.0000444496.24228.2c, URL: https://journals.lww.com/ajnonline/Fulltext/2014/03000/The_Systematic_Review__An_Overview.28.aspx?casa_token=yzXlt8721yEAAAAA:oM2-HEbo6rJ_nSHVpOQMHGC2O8Zvh6-No1H70Fmw7_Z3E23VICILvs5VlwMkMy9jOj3L2UDLsfTlAaGQ7dilr-qa.

Bastian, M., Makhortykh, M., Harambam, J., & van Drunen, M. (2020). Explanations of news personalisation across countries and media types. *Internet Policy Review*, *9*(4), 1–34.
Clark, J. D., Faris, R. M., Morrison-Westphal, R. J., Noman, H., Tilton, C. B., & Zittrain, J. L. (2017). The shifting landscape of global internet censorship.
Conway, M. (2020). Routing the extreme right: challenges for social media platforms. *The RUSI Journal*, *165*(1), 108–113.
Cotterrell, R. (1999). Transparency, mass media, ideology and community. *Journal for Cultural Research*, *3*(4), 414–426.
Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, *18*(3), 410–428. http://dx.doi.org/10.1177/1461444814543163, Publisher: SAGE Publications.
De Gregorio, G. (2020). Democratising online content moderation: A constitutional framework. *Computer Law & Security Review*, *36*, Article 105374.
European Commission, D. (2020). Proposal for a regulation of the european parliament and of the council on a single market for digital services (digital services act) and amending directive 2000/31/EC. URL: https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1608117147218&uri=COM%3A2020%3A825%3AFIN.
European Parliament, D. (2022). Texts adopted - Digital Services Act ***I - Tuesday, 5 July 2022. URL: https://www.europarl.europa.eu/doceo/document/TA-9-2022-0269_EN.html.
Facebook (2021). Transparency reports | Transparency Center. URL: https://transparency.fb.com/data/.
Flyverbom, M. (2015). Sunlight in cyberspace? On transparency as a form of ordering. *European Journal of Social Theory*, *18*(2), 168–184.
Fox, J. (2007). The uncertain relationship between transparency and accountability. *Development in Practice*, *17*(4–5), 663–671.
Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media* (Illustrated ed.). New Haven: Yale University Press.
Gillespie, T., Aufderheide, P., Carmi, E., Gerrard, Y., Gorwa, R., Matamoros-Fernández, A., Roberts, S. T., Sinnreich, A., & West, S. M. (2020). Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. *Internet Policy Review*, *9*(4), Article–number.
Github (2021). 2020 Transparency Report. URL: https://github.blog/2021-02-25-2020-transparency-report/.
Google (2021). Google Transparency Report. URL: https://transparencyreport.google.com/?hl=en.
Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), Article 2053951719897945.
Gorwa, R., & Garton Ash, T. (2020). Democratic transparency in the platform society. *Social Media and Democracy: The State of the Field, Prospects for Reform*, 286.
Greenleaf, G. (2022). Now 157 Countries: Twelve Data Privacy Laws in 2021/22. URL: https://papers.ssrn.com/abstract=4137418.
Heald, D. A. (2006). Varieties of transparency. In *Transparency: The key to better governance?: Proceedings of the British academy 135* (pp. 25–43). Oxford University Press.
Heldt, A. P. (2019). Reading between the lines and the numbers: an analysis of the first netzdg reports. *Internet Policy Review*, *8*(2).
Hood, C., & Heald, D. (2006). *Transparency in historical perspective*. Oxford University Press.
Hootsuite (2021). TikTok Hits 1 Billion Users—Faster Than Facebook (And More New Stats). URL: https://blog.hootsuite.com/simon-kemp-social-media/.
Hovyadinov, S. (2019). Toward a more meaningful transparency: Examining Twitter, google, and facebook's transparency reporting and removal practices in Russia. In *Google, and Facebook's transparency reporting and removal practices in Russia (November 30, 2019)*.
Inc, S. (2021). Snap Inc.. URL: https://snap.com/en-US/privacy/transparency.
Kosta, E., & Brewczyńska, M. (2019). Government access to user data: Towards more meaningful transparency reports. In R. Ballardini, P. Kuoppamäki, & O. Pitkänen (Eds.), *Regulating Industrial Internet Through IPR, Data Protection and Competition Law. (KLUWER LAW INT 2019)*, *Kosta & BrewczyŃSka, Government Access To User Data: Towards more Meaningful Transparency Reports*.
LinkedIn (2021). Our Transparency Report. URL: https://about.linkedin.com/transparency.
Lumen (2021). Lumen. URL: https://www.lumendatabase.org/.
MacCarthy, M. (2020). Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry. http://dx.doi.org/10.2139/ssrn.3615726, URL: https://papers.ssrn.com/abstract=3615726.
MacKinnon, R., Hickok, E., Bar, A., & Lim, H.-i. (2015). Fostering Freedom Online: The Role of Internet Intermediaries. *Other Publications from the Center for Global Communication Studies*, URL: https://repository.upenn.edu/cgcs_publications/21.
Makhortykh, M., Urman, A., Münch, F. V., Heldt, A., Dreyer, S., & Kettemann, M. C. (2022). Not all who are bots are evil: A cross-platform analysis of automated agent governance. *New Media & Society*, *24*(4), 964–981. http://dx.doi.org/10.1177/14614448221079035, Publisher: SAGE Publications.
Makhortykh, M., Urman, A., & Wijermars, M. (2022). A story of (non)compliance, bias, and conspiracies: How Google and Yandex represented Smart Voting during the 2021 parliamentary elections in Russia. *Harvard Kennedy School Misinformation Review*, http://dx.doi.org/10.37016/mr-2020-94, URL: https://misinforeview.hks.harvard.edu/?p=8944.
Microsoft (2021a). Content Removal Request Report | Microsoft CSR. URL: https://www.microsoft.com/en-us/corporate-responsibility/crrr,
Microsoft (2021b). Digital Safety Content Report | Microsoft CSR. URL: https://www.microsoft.com/en-us/corporate-responsibility/digital-safety-content-report?rtc=1&activetab=pivot_1:primaryr3.
Microsoft (2021c). Law Enforcement Request Report | Microsoft CSR. URL: https://www.microsoft.com/en-us/corporate-responsibility/lerr,
O'Regan, T., & Li, L. (2019). Recalibrating China in a time of platforms. In *Digital transactions in Asia* (pp. 63–82). Routledge.
Parsons, C. (2019). The (In)effectiveness of Voluntarily Produced Transparency Reports. *Business & Society*, *58*(1), 103–131. http://dx.doi.org/10.1177/0007650317717957, Publisher: SAGE Publications Inc.
PCMAG (2021). Definition of Big Tech. URL: https://www.pcmag.com/encyclopedia/term/big-tech.
Pinterest (2021). Transparency report. URL: https://policy.pinterest.com/en/transparency-report.
Post, W. (2021). Human rights advocates decry Apple, Google decision to pull Navalny app as Russia voting begins. *Washington Post*, URL: https://www.washingtonpost.com/business/2021/09/17/navalny-google-apple-app-russia/.
Principles, S. C. (2021). Santa Clara Principles on Transparency and Accountability in Content Moderation. URL: https://santaclaraprinciples.org/images/santa-clara-OG.png.
Reddit (2021). Transparency Report 2020 - Reddit. URL: https://www.redditinc.com/policies/transparency-report-2020-1.
Sivetc, L., & Wijermars, M. (2021). The vulnerabilities of trusted notifier-models in Russia: The case of netoscope. *Media and Communication*, *9*(4), 27–38.
Sneed, M. (2020). The key to regulating facebook and data collection companies is transparency. *Album LJ Science & Technology*, *30*, 109.
Stecklow, S., & Dastin, J. (2021). Special Report: Amazon partnered with China propaganda arm. *Reuters*, URL: https://www.reuters.com/world/china/amazon-partnered-with-china-propaganda-arm-win-beijings-favor-document-shows-2021-12-17/.
Stier, S., Bleier, A., Lietz, H., & Strohmaier, M. (2018). Election campaigning on social media: Politicians, audiences, and the mediation of political communication on facebook and Twitter. *Political Communication*, *35*(1), 50–74.
Suzor, N. P., West, S. M., Quodling, A., & York, J. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication*, *13*, 18.
TikTok (2021). Tiktok transparency report. URL: https://www.tiktok.com/safety/resources/transparency-report-2020-2?lang=en.
Twitch (2021). Transparency Report. URL: https://safety.twitch.tv/s/article/Transparency-Reports?language=en_US#2H12021TransparencyReport.
Twitter (2021). Twitter Transparency Center. URL: https://transparency.twitter.com/en.html.
Urman, A. (2020). Context matters: political polarization on Twitter from a comparative perspective. *Media, Culture & Society*, *42*(6), 857–879.

Urman, A., Ho, J. C.-t., & Katz, S. (2021). Analyzing protest mobilization on telegram: The case of 2019 Anti-Extradition Bill movement in Hong Kong. *PLoS One*, *16*(10), Article e0256675.

van Drunen, M. Z., Helberger, N., & Bastian, M. (2019). Know your algorithm: what media organizations need to explain to their users about news personalization. *International Data Privacy Law*, *9*(4), 220–235. http://dx.doi.org/10.1093/idpl/ipz011.

Wagner, B., Rozgonyi, K., Sekwenz, M.-T., Cobbe, J., & Singh, J. (2020). Regulating transparency? Facebook, Twitter and the german network enforcement act. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 261–271).

Wu, K. (2016). YouTube marketing: Legality of sponsorship and endorsements in advertising. *JL Business & Ethics*, *22*, 59.

Zalnieriute, M. (2021). *"Transparency-washing" in the digital age: A corporate agenda of procedural fetishism: SSRN scholarly paper ID 3805492*, Rochester, NY: Social Science Research Network, URL: https://papers.ssrn.com/abstract=3805492.

Dr. **Aleksandra Urman** is a postdoctoral researcher at Social Computing Group, University of Zurich. Aleksandra's research interests include political communication on social media, algorithmic biases, and computational research methods.

Dr. **Mykola Makhortykh** is a postdoctoral researcher at the University of Bern, where he studies information behavior in online environments. Before moving to Bern, Mykola defended his Ph.D. dissertation at the University of Amsterdam on the relationship between digital platforms and war remembrance in Eastern Europe and worked as a postdoctoral researcher in Data Science at the Amsterdam School of Communication Research, where he investigated the effects of algorithmic biases on digital news consumption.