Transparency Reporting: The UK Regulatory Perspective

Anna-Sophie Harling, Declan Henesy, and Eleanor Simmance

1 Introduction

Around the globe, there are growing calls from politicians, policymakers, academics, and civil society for greater transparency from online platforms.[1] But what this means in practice is not always clear.[2] Some proposals focus narrowly on particular companies, whereas others are very broad in scope. And there are widely varying views about what constitutes "meaningful information" and how to balance different objectives. There is a risk of advocating for greater transparency for transparency's sake, without having a clear purpose of what it is and what it's for.

Many online platforms already publish transparency reports. Google published its first transparency report in 2010, Twitter in 2012, and Facebook in 2013. More recently, medium-sized services have followed suit, including Nextdoor, Discord, and Yubo, as well as adult sites, such as PornHub.[3] However, reporting to date has been conducted largely on a voluntary basis, with companies choosing how they undertake reporting, what information they disclose, when, and in what format. These reports focus heavily on metrics rather than process and provide only a partial account of what's happening inside companies and across the platforms they operate.

We work at Ofcom, the United Kingdom's independent communications regulator, which is set to take on regulatory powers to protect UK users from harms occurring on user-to-user services and search services. The Online Safety Bill, introduced into Parliament in March 2022, is presently making its way through Parliament. The bill includes provisions conferring powers on Ofcom to establish a mandatory platform transparency reporting regime.[4]

[1] Aspen Institute. 2021. "Final Report: Commission on Information Disorder"; Council of the European Union. 2019. "EU Introduces Transparency Obligations for Online Platforms"; Dawes, Dame M. 2021."In News We Trust: Keeping Faith in the Future of Media." Ofcom, October 7, 2021; douek, evelyn. 2022."Content Moderation as Systems Thinking." SSRN Scholarly Paper; Integrity Institute. 2021."Metrics & Transparency: Data and Datasets to Track Harms, Design, and Process on Social Media Platforms"; Myers, Steven Lee. 2022. "Obama Calls for More Regulatory Oversight of Social Media Giants." The New York Times, April 21, 2022, sec. Technology.

[2] Keller, Daphne. 2021. "Some Humility About Transparency." The Center for Internet and Society; Tworek, Heidi, and Alicia Wanless. 2022. "Time for Transparency From Digital Platforms, But What Does That Really Mean?" Lawfare. January 20, 2022; MacCarthy, Mark. 2021. "How Online Platform Transparency Can Improve Content Moderation and Algorithmic Performance." Brookings. February 17, 2021; Sorensen, Kiki. 2021. "What's Wrong with Transparency Reporting (and How to Fix It)." ADL. October 12, 2021.

[3] Nextdoor. 2022. "Nextdoor Transparency Report"; Yubo. 2022. "Transparency Report 2021"; Cole, Samantha. 2021. "Pornhub Just Released Its First Transparency Report." Vice; Discord. 2019. Discord Transparency Report 2019.

[4] As set out in Ofcom's Roadmap to Regulation, published in July 2022, we expect the transparency reporting regime to come into effect in 2024.

doi:10.54501/jots.v1i5.108

The industry in scope of the UK's online safety regime is entirely different in nature, pace, and scale from the sectors Ofcom has traditionally regulated, such as telecoms or postal services. It is an industry that has not to date faced comprehensive regulatory oversight of its trust and safety practices, depriving the public of systematic insight into how decisions are made, how their products are designed, and how these affect users and societies around the world. To meet the challenges of regulating a wide-ranging, rapidly changing set of online platforms, Ofcom will have new regulatory tools at its disposal, including issuing mandatory transparency notices to platforms and publishing transparency reports.

We believe there is a benefit in rethinking the approach to transparency reporting. We begin this paper by exploring our main objectives for transparency reporting, and subsequently explore the challenges, risks, and limitations of the ways platforms currently publish information and metrics. We also touch on the value of international alignment, which can help us create an efficient and impactful transparency regime that builds understanding about online services and prioritizes users' safety. We will continue to engage with the public on our approach to transparency reporting and welcome feedback on this paper in advance of formal consultation.

2 Transparency Reporting Objectives

The main objective of the Online Safety Bill is to improve the safety of users online. As we set out in the roadmap we published in July 2022, the Bill gives us powers to achieve this through both increased understanding by the public of harms and measures on platforms, as well as systemic changes to platform design, safety measures, and governance.[5] Transparency lies at the core of this regulation.

Under the Online Safety Bill, Ofcom will be required to issue transparency notices to a subset of in-scope platforms; these may be tailored to each particular service, specifying the information and data different platforms must publish, the methodology used, and the format in which the information is gathered and published.[6]

Ofcom will also be obligated to publish its own transparency reports each year, based on the information published in platforms' transparency reports.[7] These powers are to be underpinned by a robust framework for enforcement in the event of noncompliance. In the case of user-to-user services, for example, transparency reports produced under the Online Safety Bill may include information about the incidence and dissemination of illegal or harmful content and the number of users who have encountered this content. Ofcom might also require platforms to explain in their transparency reports how they enforce their policies and community guidelines, publish information about user reporting systems and user empowerment tools, or disclose details about content moderation technologies and user identity verification. Other areas of focus might be corporate governance structure and decision-making, risk assessment outcomes, and internal key performance indicators across teams. As described above, the transparency framework under the current bill requires two separate but related outputs: platform transparency reports and Ofcom transparency reports. These could

[5] Ofcom. 2022. "Online Safety: Ofcom's Roadmap to Regulation."

[6] When determining what information to require in a notice, Ofcom will have to balance a number of different factors, taking into account the type of service, its functionalities, the number of users, and the capacity of the provider in question (among other things).

[7] These would contain a summary of the conclusions drawn from platforms' transparency reports, including any cross-industry patterns or trends, a summary of measures Ofcom considers to be good industry practice, and any other information from the platform transparency reports that Ofcom deems appropriate to include.

achieve the various objectives discussed below.

Empower the public: Both platform and Ofcom transparency reports could be relevant to not only users of a service, but also a range of non-governmental, commercial, and public bodies. For example, transparency information could help:

- Online services to self-correct where issues are revealed or to adopt industry best-practice;
- Journalists or researchers to develop public understanding of online harms and society's broader information environment;
- Civil society organizations to research online harms and potential mitigations;
- Investors or shareholders to direct capital flows to responsible companies; and
- Advertisers, payment providers, etc. to assess commercial relationships with online services.

In different ways, each of these audiences can use the information in both reports to better understand how technology companies operate and hold them accountable to make meaningful, systemic changes. We will consider how best to meet the needs of different audiences as we develop our thinking in more detail.

Test and communicate the effectiveness of the regime: Platform transparency reports may help Ofcom fill key evidence gaps, test whether measures are delivering the right outcomes, and inform our regulatory strategy. Ofcom is also required to use the information published in platform transparency reports in its own transparency reports.

Encourage proactive, meaningful action from platforms: The publication of key information can help drive change in regulated services.[8] Online platforms, in particular, care deeply about avoiding negative press and placating advertisers, who don't want

[8] There remain challenges in testing empirically the role of transparency in reducing harms, as these effects may take time to feed through and are difficult to isolate alongside other confounding factors (e.g., other regulation that might be introduced at the same time). In addition, regulation for online services is still in its infancy, so there is scant empirical evidence on the effectiveness of transparency in the digital space. Nevertheless, some academic literature has looked at information disclosures in other settings and has provided some evidence that information disclosures may contribute to harm reduction efforts under Ofcom's online safety regime. These settings include both voluntary and mandatory transparency measures relating to environmental harms, corporate social responsibility (CSR), and food hygiene. The empirical evidence reviewed suggests that firms can respond to transparency by reducing harms that are exposed by meaningful disclosures. The impact of environmental disclosures on harm is one of the contexts that has received particular focus from transparency researchers. For instance, Bennear and Olmstead (2008) examine firms' responses to mandatory water safety disclosures. The authors' findings are suggestive of disclosures reducing annual drinking water violations. Delmas et al. (2010) examine the impact of mandatory information disclosures on the fuel mix used by firms generating and supplying electricity. Their findings suggest that as the proportion of a firm's sales subject to mandatory disclosures increases, the average proportion of generation attributable to fossil fuels drops. Finally, Chen et al. (2018) examine the impact of a 2008 mandate requiring Chinese firms to disclose information on their environmental activities. The researchers' findings suggest that firms in cities most impacted by the disclosure requirements experienced a decrease in industrial wastewater and sulfur dioxide (SO_2) emissions. Transparency research related to CSR disclosures has also attracted much attention in the literature. Fiechter et al. (2019), for instance, consider whether firms within scope of a new EU transparency directive increase their CSR activities following information releases. The researchers' findings suggest that firms subject to the regulation on average increase their CSR activities relative to a sample of US-based companies who act as controls. They also find suggestive evidence that improvements were more pronounced for firms facing larger increases in mandated CSR disclosures. Evidence on the effectiveness of transparency in the environmental and CSR contexts, therefore, broadly suggests that firms respond to mandated disclosures by reducing harm.

their brands associated with harmful content.[9] Revelations about online platforms failing to prioritize user safety can have immediate impacts on user numbers, advertising spend, and share prices. Targeted transparency requirements will be a major tool for driving behavior change under Ofcom's online safety regime. Ofcom's ability to protect users from harm online therefore hinges on an effective transparency regime.

3 Rethinking Metrics

With some exceptions, transparency reports produced by platforms have generally focused on content moderation metrics and government requests around user data and records.[10] Larger platforms tend to report more—and more detailed—metrics and accompanying commentary or materials, as they generally have larger Trust & Safety teams and budgets to put behind their transparency reports. The metrics reported by the largest platforms vary from number of content removals and user reports to estimates of the prevalence of violative content. As Ofcom prepares to implement its mandatory transparency regime, we are carefully considering the challenges, risks, and limitations around such metrics.

To start with, the content removal metrics currently reported by platforms provide some insight, but have their limitations. Consider the metric "140,000 pieces of hate speech removed in Q1." If this figure reportedly goes up in Q2, does that mean there was more hate speech on the platform than in Q1? Does it mean that the systems in place to identify this content became more effective? Does it mean that the platform changed its definition of hate speech in Q2, resulting in a greater number of pieces of content violating its rules? What was the impact of major international, national, or local events on the amount of hateful content uploaded or resurfaced by users in Q2?

In recent years several major platforms, such as Facebook, Snapchat, and YouTube, have converged on the view that exposure-based metrics (which estimate how many times users see violative content) should play a central role in discussions about harm reduction and platform accountability.[11] However, exposure metrics say little about who viewed content and nothing about the impact of that exposure in terms of actual personal (e.g., psychological, financial) or social harm. The link between exposure

[9] For example, following internal changes at Twitter made by Elon Musk in November 2022,GroupM designated the platform as a "high risk" media buy. GroupM called on Twitter to return to baseline levels of harmful content, staff up its Trust & Safety team, demonstrate its commitment to effective content moderation and enforcement of policies, and improve transparency around brand and user safety.

[10] Such figures have served as the baseline for recent efforts to create standards for consistent and comparable data across platforms that wouldn't be linked to platforms' terms of service, which are liable to change. Notable examples include the government-led OECD Voluntary Transparency Reporting Framework on terrorist content, or the investor-led Value Reporting Foundation's SASB Standard on Content Moderation on Internet Platforms (Content Moderation on Internet Platforms - SASB). Relatedly, corporate benchmarks such as Ranking Digital Rights or the Global Child Forum Benchmark seek to compare companies' policies and performances relevant to each other through scoresheets or leaderboards. The objective here is to inform customers and investors and to reward good practice by naming and acclaiming or shaming. While this approach has effectively spurred pro-social competition in some cases, there are also fundamental challenges in seeking to create comparable metrics across different services, business sizes, target user base, etc. These standards and benchmarks have been shaped by (and continue to influence) parallel legislation around transparency and corporate reporting, such as the 2014 EU Non-Financial Reporting Directive or the new 2021 Corporate Sustainability Directive. Other digital regulators have also sought to build on these trends through their policy agendas. For example, Australia's eSafety Commissioner produces a number of resources for investors and venture capitalists, including an investment checklist, model clauses for due diligence arrangements, and an assessment tool for startups.

[11] Meta. n.d. "Prevalence"; YouTube. n.d. "Violative View Rate"; Snap. 2022. "Transparency Report: Snap's Violative View Rate." Other metrics about views of violative content include the amount of viewing removed or content received, e.g., Pinterest. 2022. "Transparency Report: Reach of Deactivated Pins"; TikTok. 2022. "Community Guidelines Enforcement Report: Removal Before Any Views."

and harm is likely to be highly complex and to depend on—among other things—the vulnerability of those accessing content, the frequency and content of exposure, etc.

Another challenge is that most reported exposure metrics aggregate data for whole user bases over extended periods of time. As a regulator, we may be more interested in the concentrated areas of harm that exist across platforms, such as the 5% of users under 18 who encounter an overwhelming amount of terrorist or suicide content. Even if the overall levels of content violating a platform's policies are low (or aggregated exposure to such content is low), there could still be a risk of harm if users with particular vulnerabilities or characteristics, such as children, are more likely than average to be exposed on a repeated basis.

Another contentious issue when considering the value of metrics is the potential for standardization across the industry. While it is important for providers to report information and metrics that are relevant to their services, some form of comparability would, in theory, allow researchers, policymakers, and other interested parties to compare online safety efforts and reveal how well systems and processes are working across the sector. When considering this, we expect to focus on what information is useful to Ofcom and the public, as well as risks associated with standardization, such as stifled innovation and disproportionate burdens on smaller platforms. Services and their policies also vary significantly, meaning the information they record may be difficult or unhelpful to compare. The definition of user-to-user service in the Online Safety Bill is broad and can encompass not just social media platforms and search engines, but also dating apps, gaming providers, online marketplaces, and more. Each of these services has varying functionalities—e.g., ephemeral content, livestreaming, audio chat, algorithmic feeds, video autoplay—that could render standardized metrics meaningless.

One question we are considering is whether it would be beneficial to establish a baseline set of metrics that all services can report on, and then build on this by requesting service-specific information.[12] This baseline information will be informed through our online safety call for evidence and consultation, as well as other relevant sources of information. For example, we are continually carrying out analyses of current platform transparency reports and civil society resources and research, such as the Santa Clara Principles on Transparency and Accountability around content moderation, Ranking Digital Rights' Corporate Accountability Index, and resources created by the Integrity Institute around transparency.

If the metrics currently reported tell us one thing, it is the sheer scale of content moderation happening in real time across the platforms that will soon fall in scope of the online safety regime. But if we want to trigger lasting changes across the sector, Ofcom cannot simply regulate content moderation processes. Online safety regulation should seek to effect systemic changes to platform design, architecture, and corporate governance. Indeed, a wide range of mitigations is available to online services, with varying degrees of effectiveness, relevance to different harms, and impact on user freedoms. We will consider how transparency reporting can go beyond content moderation to address the different ways that services protect their users from online harms and highlight good and bad practice, all the while keeping in mind the potential risks around arming bad actors with information on how to circumvent safety systems.

[12] This may require Ofcom to carefully specify the methods platforms use to gather and report this data to ensure comparability between platforms and for consistent collection year on year. A baseline set of metrics may also make the most sense for illegal content and content that is harmful to children, as the rules here are consistent across all platforms, though further exploration is needed.

Finally, the transparency reporting duties in the Online Safety Bill should encompass qualitative information. Details of content moderation procedures; implementation of safety measures such as age verification and hashing; oversight and governance arrangements; and reporting, flagging, and appeals processes are examples of information that provides meaningful insight into how platforms operate. As we develop our transparency notices, we will consider the full range of information to request from platforms. Particularly in the first years of the regime, we anticipate requiring disclosure of both qualitative information and select quantitative data and metrics, for the purpose of assessing effectiveness and impact over time. The Online Safety Bill requires services to take into account the risk of illegal harms and harms to children when developing their products and services. Crucially, our transparency powers will as currently drafted give us the ability to require platforms to publish information relating to their risk assessments, including risk assessments carried out when the service is being designed, when updates to the service are being considered, and while the service is in operation. This would allow for public accountability and scrutiny by civil society groups and academics and would enable proactive regulatory intervention when products are being developed.

4 The Importance of International Alignment

Alignment with international legislation remains an important consideration for Ofcom's transparency regime. To what extent is there scope to create efficiencies for platforms and align transparency requirements with other regimes such as the Digital Services Act (DSA)? To what extent do we want to pursue a different approach to transparency reporting?

We recognize the potential value of convergence in this area. A global approach to transparency reporting could minimize burdens on large and medium-sized platforms in scope of numerous regulatory jurisdictions. However, the UK and EU legislative frameworks do vary, and this may enable us to take a more dynamic and flexible approach. For example, the Online Safety Bill presently gives Ofcom the ability to send tailored transparency notices to services, which can take into account the nuances between different types of platforms and develop over time to adapt to new technologies and harms. In contrast, the DSA, as well as legislative proposals currently being discussed in the US, sets out standardized transparency requirements on platforms.[13] The UK's more dynamic approach could therefore be integral to spurring innovation and encouraging best practice in the field of transparency.

Policy changes in one country can drive product changes at a global level. This was seen with the introduction of the UK's Age-Appropriate Design Code,[14] which influenced stronger child safety measures globally.[15] However, there is also precedent that services might choose to treat their UK transparency reporting requirements as separate from the rest of their transparency reporting duties, treating them as an add-on rather than raising international standards more broadly. Under Germany's NetzDG, social media platforms are required to create and publish biannual reports that detail

[13] The DSA gives the European Commission the power to request specific information from services in the context of investigations. Ofcom's transparency notices, in contrast, require platforms to publish specific information for public consumption and are not issued as part of investigations.

[14] The Age-Appropriate Design Code was introduced by the UK data protection regulator, the Information Commissioner's Office (ICO), in September 2021: ICO. 2021. "Age appropriate design: a code of practice for online services."

[15] BBC News. 2021. "Children's internet code: What is it and how will it work?"; 5Rights Foundation. 2021. "TikTok announcement shows impact of Children's Code"; TikTok. 2021. "Strengthening privacy and safety for youth on TikTok"; Instagram. 2022. "Introducing New Ways to Verify Age on Instagram."

their handling of illegal content.[16] Yet providers such as Google, Meta, and Twitch have chosen to produce standalone NetzDG transparency reports focusing on specific transparency efforts in Germany, rather than adapt their core operations to apply transparency standards more widely. Fostering alignment with international regulation will help to mitigate these risks. Ofcom expects to consider whether it is appropriate to align transparency requirements under the Online Safety Bill with other international approaches including, for example, the approach under the DSA. We could then build on these baseline requirements to require more in-depth, meaningful information from platforms about the ways their products work to keep users safe.

This may allow Ofcom to take a more targeted approach while maintaining a set of core indicators to permit comparative analysis over time and across jurisdictions. Product changes can happen at a global level, meaning that a successful transparency regime might nudge platforms to make systemic changes that impact users around the world.

5 Transparency Matters

Transparency will be a powerful and essential tool in our regulatory arsenal. As the future online safety regulator, we plan to think long and hard about the numerous challenges and trade-offs associated with mandatory transparency reporting. Any requirements we place on platforms could have far-reaching consequences for internet users beyond UK borders. This presents great risk, but even greater opportunity. A carefully designed transparency regime could transform Ofcom's ability to hold platforms accountable and fundamentally change the way the industry prioritizes the safety of its users.

[16] The NetzDG law requires social media platforms to establish a transparent procedure for dealing with complaints about illegal content, which is subject to a reporting and documentation obligation. Platforms should check complaints immediately, delete "obviously illegal" content within 24 hours, and delete and block access to any illegal content within seven days after checking. In addition, providers must submit a six-monthly report on complaints received and how they have been dealt with.

Authors

Anna-Sophie Harling is a Principal at Ofcom, leading work on online safety transparency reporting. (ash@ofcom.org.uk)

Declan Henesy is a Policy Manager in the Online Safety Policy team at Ofcom, working on transparency reporting and content moderation policy.

Eleanor Simmance is a Senior Analyst in the Research & Intelligence team at Ofcom, working on transparency and measurement.

Acknowledgements

The authors would like to thank colleagues at Ofcom working across policy, research, economics, and legal issues relating to online safety and transparency. We would also like to thank our counterparts internationally and those working in trust and safety at platforms, civil society organizations, research organizations, and so on, whose contributions through consultation responses and in conversations with us are invaluable to our understanding of the sector as we prepare to take on our new powers.

Keywords

Transparency; reporting; regulation; metrics; Ofcom; UK