ONLINE SAFETY ACT
NETWORK

RESPONSE TO OFCOM CONSULTATION ON DRAFT TRANSPARENCY GUIDANCE

Summary

1. We welcome the opportunity to respond to Ofcom's consultation on its draft transparency guidance. This response reflects discussions with and contributions from members of the Online Safety Act Network, many of whom have a strong interest in the effective implementation of the transparency measures in the Online Safety Act. Individual organisations within our network will be submitting their own responses to the consultation so this response is not intended to speak on their, or on any other organisations', behalf.

2. We have submitted material from this response via the Ofcom consultation proforma, where it is relevant to the specific consultation questions, but would request that this response is also considered in its totality given that it does not neatly map onto all of those questions. This is particularly important given that the consultation proforma specifically asks for industry responses without a mention of seeking a similar input from civil society or organisations representing users[1]. This is despite Ofcom setting out in the consultation documents - in a way that is welcome - that the two main outcomes they wish to deliver from transparency reporting are focused on users: improved "safety outcomes for UK users on their service" and increasing "users' understanding of regulated services, enabling them to make informed choices about how they spend their time online." More explicitly framing the consultation questions in these terms - eg will these proposals deliver the outcomes for users we have identified? - and making it clear that a wide range of stakeholders might have views on that would have been advisable, instead of prioritising feedback from industry on whether they are happy with the proposals insofar as they are required to comply with them.

Our response

3. We focus our response on a number of key issues around which we have concerns, or where we have suggestions for Ofcom in terms of ensuring that their transparency tools are used in the most effective way possible. We note that the timescales for the issuing of the first transparency

[1] "We welcome input from industry on the areas listed below. We encourage stakeholders to respond with feedback." (Consultation response form, page 1.)

notices is contingent on the DSIT Secretary of State's decision on Ofcom's advice on categorisation thresholds, without which the register of categorised services cannot be published. While it is unfortunate that there is this delay, we still strongly believe that Ofcom's advice to the Secretary of State on categorisation was flawed (see our analysis here) and that, should a different approach be taken to allow for the inclusion of small but risky services within category 1, the transparency measures will go much further to deliver Ofcom's stated aims when also applied to those smaller, riskier services.

4. We welcome the clarity of the proposals set out by Ofcom and its expectations that these transparency measures will provide stronger safety governance, services designed and operated with safety in mind, greater choice for users and greater transparency about the safety measures used by services which should lead to greater user trust in those services. It is useful also to understand how Ofcom views the role of transparency measures in driving up standards, not least because large corporations and companies are responsive to legal, commercial, reputational risks arising from the disclosure of information while shining a light on best practice can lead to overall improvements across a sector.

5. We note that the publication of this information will also be useful to researchers, users, civil society, advertisers, investors, shareholders etc. We would however question whether it is the role of civil society organisations - many of them under-resourced - to, as Ofcom suggests, "create resources to help industry implement changes to their systems and processes"; similarly, in relation to the outcome on "increasing users' understanding of regulated services and enabling them to make informed choices about how they spend their time online", we note that the regulator is looking to civil society's support: "Civil society can communicate information relevant to those specific groups. This will support our efforts …".

6. We hope that in return for passing on these expectations that civil society will play their role in improving the safety efforts of the companies regulated by Ofcom, Ofcom will also take into account the suggestions from civil society organisations fed in through this consultation as to the role they could play in shaping the asks of those companies prior to the issuing of the first transparency report. We would also hope in this regard that the specific recommendations from civil society on what good transparency reporting might look like - and the lessons that can be learned from the failure of voluntary transparency reporting to deliver true accountability - are taken on board. This will go a long way to ensuring that the first round of transparency reporting meets their expectations as much as it meets industry's and that, as a result, the resources Ofcom wishes them to provide when the reports are available are as effective in shifting the dial on user safety for the users they represent as they can be.

Ofcom's ambitions

7. While we welcome the overall approach being taken by Ofcom and agree with them that transparency reporting is a powerful tool, we raise here a few areas where we are concerned

that the regulator may be limiting its scope and constraining its powers unnecessarily. We have written before about the fact that whatever baseline Ofcom sets in its first OSA codes or guidance will remain in place for a number of years before further iterations are produced. We are also very well aware that industry responses to this consultation - and lobbying during the ahead of the first round of engagement - risks watering down that baseline before a full implementation cycle has taken place. Consequently it is vital that Ofcom starts from the strongest position possible - within the scope that the OSA provides it. With that in mind, there are a number of areas where we feel that Ofcom could go further.

8. We would like to see a more explicit commitment to making harm reduction an intended outcome of the transparency process and for commitments from Ofcom that the metrics they will use in their transparency notices measure user experiences of harmful content in order to provide information on the effectiveness of otherwise of a providers' safety measures.

9. More generally, in the discussion on the principles identified and discussed in its draft guidance, Ofcom talks about "relevance" and "appropriateness"; it then goes on to talk about proportionality: "*Another key principle that we set out in the draft guidance is proportionality. We will always take steps to ensure that the requests for information in our notices go no further than is necessary to give effect to our policy objectives*. (para 3.20) While we appreciate that any regulator cannot and should not be engaged in a widespread fishing exercise when seeking information from regulated services - and that Ofcom's policy objectives are broad - we have also raised concerns in response to Ofcom's earlier consultations that proportionality is generally interpreted by them as relating to costs, eg not imposing extra financial or resourcing costs on companies. Some of the policy objectives are very weighty indeed.

10. We appreciate that proportionality is very relevant to SMEs across the broad set of duties that apply to all regulated services but, as things stand, it is only categorised services who will be required to produce transparency reports; the Act itself already provides these limitations on the grounds of proportionality. On top of this, Ofcom has - in its advice to the Secretary of State - limited "category 1" to the largest user-to-user services. If "proportionality" therefore means that these largest, multinational companies can challenge Ofcom's requests based on cost (as para 3.23 in the guidance suggests, see below) then Ofcom is further limiting the potential extent and impact of the transparency requirements, unnecessarily constraining itself and risking the fact that the information it receives back will not meet the expectations or outcomes it has set itself.

11. We set out in the next section some further specifics contained in the guidance where we feel Ofcom has unnecessarily limited its approach without first testing whether a more comprehensive, expansive approach would be practical or seeking input from stakeholders other than industry as to the impact this might have.

<u>The guidance</u>

12. The draft guidance sets out that Ofcom can require information to be included in the transparency reports based on the list provided in Schedule 8 of the Online Safety Act. This includes a list of specific proposed topics plus “any other measures taken or in use by a provider which relate to online safety matters”. Ofcom’s “online safety functions” are also much broader than enforcing the Online Safety Act duties and cover broader functions. We would welcome it if Ofcom can make this clear in the final version of its guidance which otherwise, as we set out below, appears to limits Ofcom’s approach in a way that is to the benefit of the regulated providers rather than to the needs of the stakeholders who Ofcom hopes will gain from the publication of the transparency information.

13. We understand that Ofcom cannot ask for any or all information and, for transparency reporting to be useful, the regulator needs to limit the information it requests, to be specific to the service in question and to cover the topics and areas of work where they think they can drive positive change. We have concerns, however, that this potentially hinders the development of baseline comparisons in relation to user safety.

14. Yet, Ofcom’s narrative in the guidance includes the following statements without providing further context as to what information might fall outside the scope of its requests, nor how it has come to the decision as to what is within or outside scope. It is implied that these judgements are incontrovertible when in fact they are just that: judgements:

    - “We will take steps to ensure that the requirements are not unduly onerous.”
    - Their understanding of risks will “narrow the topics of information that we require providers to report on”.
    - “we will be particularly mindful of the category within which a service falls and the duties that apply to the service as a consequence” - this in particular is problematic, we believe, in contrast to the online safety matters within Schedule 8.
    - “In our notices, we will not ask services to publish information about duties that it is not required to comply with” - does this preclude asking for information to support media literacy, for example?
    - Ofcom will “allow the opportunity for representations about the likely time, cost and effort to give effect to the proposals” (para 3.23)

15. We are concerned that this narrowing of intent will not give Ofcom the breadth of information it needs in order to either inform subsequent iterations of the codes of practice or to shine a light on where are gaps in good practice across services. There is a significant information asymmetry between the major social media platforms and regulators, researchers and others who wish to hold them to account and, while Ofcom hopes that their approach will lead to better outcomes, the tone of the document suggests that the power remains in the hands of industry with regard to what happens next.

16. There is also plenty of emerging evidence - in particular from the many legal actions in the US - that the transparency reporting provided on a voluntary basis by many of the major platforms is often not meaningful, and in many cases it is misleading. Voluntary reporting by platforms has allowed them to set the terms of the aspects of their service they wish to report on, decide on the metrics and also decide on the contextual information provided to interpret those metrics. Large headline figures indicating big volumes of actions in relation to the moderation of particular categories of content are meaningless without understanding the overall volume of that sort of content that is passing over a service. Platforms rarely provide detail on the sequence of actions that might be taken in relation to a particular harm or a particular type of content but provide metrics in isolation that only give a partial picture.

17. More significantly for Ofcom's work in setting the terms of transparency reporting under the OSA, concerns have arisen in recent years over whether the available metrics in transparency reports even accurately relate to actions taken by the company according to their various policies and terms of services; and the vagueness of what the "actions" taken are when they are reported as taking place. For example Meta's transparency reports do not specify what their "actions" are but could cover a number of responses such as pop-up warnings, suspended accounts, deleted accounts or data reported to authorities; the aggregate (big) number may sound impressive, but it obscures both what the actions were and whether they were effective.

18. TikTok has recently started following Meta's approach, citing the percentage of "actions" taken relative to actions taken because humans report problems. Presenting numbers of actions in this way leads people to assume it means the platform is removing the vast majority of the violating content on its platform. In fact, it does not actually say that. We urge Ofcom to press for more analysis from all the major platforms on the estimated extent of violating content, and the overall percentage they believe they remove.

19. We have included in the annex to this document a number of examples to provide some evidence in this area for Ofcom to consider as they gear up to issuing their first transparency notices. This includes:
    a. Inaccurate metrics on complaints and their resolution: the Meta Oversight Board has identified instances where reports on posts were closed by automated systems as they did not receive a human review within 48 hours; these will have been recorded as "resolved" though no action was taken. (See for example, here.)
    b. Misleading information on "anonymous" accounts: claims provided by Twitter to the media, in the wake of the racist abuse of black England footballers at the Euros in 2021, that 99% of the accounts involved were not anonymous were subsequently disproved.
    c. Misleading results and discrepancies uncovered by researchers accessing the TikTok API under the EU Digital Services Act;
    d. Meta's quarterly transparency reports stating a higher CSAM detection rate;
    e. The limits of Snap's transparency reporting on both self-harm and suicide and on the promotion of Fentanyl to teenagers via their platform; Snap is also the subject of a

recent lawsuit brought by the New Mexico Attorney-General, with recently unredacted material illuminating the discrepancy between Snap's public statements of harm on its platform and the information held within the company.

20. In light of the recent investigations into, and lawsuits against, Snap - and documentation from previous US lawsuits and whisteblower revelations, which we have collated on our website - we would recommend that Ofcom also consider in its guidance a requirement that regulated platforms have a whistleblower policy and report on activities in respect of any incidences of whistleblowing and actions taken as a result within the timeframe covered by transparency report.

Industry engagement and representations

21. In terms of the process of engagement that Ofcom describes, we have specific concerns about the imbalance between Ofcom's approach and the might of industry to derail it - again we are talking here about the largest companies, as per Ofcom's advice, which is a tiny proportion of the overall number of services that fall into scope of the legislation. Ofcom proposes to introduce a step - that is not required by the legislation - to discuss the draft transparency notice with companies and give them the opportunity to make representations on whether the request is feasible. We set out the full text of the proposal here:

> "Each year Ofcom will share with providers a draft transparency notice (or, where the provider provides more than one categorised service, drafts of the transparency notices), containing the information Ofcom proposes to require services to produce in their transparency reports. This will offer the opportunity for providers to make written representations on the proposed information to be produced within the report before the notice is formally issued. Among other things, this process is intended to ensure the requests are clear, targeted and proportionate to the technical capabilities and capacity of the provider." (para 4.13)
>
> "Ofcom will provide an opportunity during the draft notice process to the provider to present any concerns arising out of such publication, including around the confidentiality of the information, and will seek to take this into account when reaching a decision. When considering any concerns, we will typically have to balance the provider's concerns around publication, including possible harm to legitimate business interests, against the extent to which publication of the information is necessary to exercise our functions around transparency. In some cases, we may consider alternatives to our requested information, where appropriate, if the alternative information is sufficient, reasonable and meets our aims." (para 4.14)

22. We understand from discussions with Ofcom that this step is very much to ensure that this process is intended to ensure the requests to providers are clear, relevant and proportionate to

the technical capacity and capabilities of the providers and it is an opportunity for Ofcom to ensure that services understand what they are being asked for and why. To avoid any perception that these discussion on the draft notices have led to a dilution of the requests to providers, and indeed to ensure full transparency, we strongly recommend that Ofcom commits to publishing both the draft transparency notice and the final draft transparency notice to enable third parties to understand which information may not have been provided as a result of providers' representations. Subject to commercial sensitivities, we would also recommend that the full text of the providers' representations to Ofcom that led to the amended final information request also be published for full transparency to understand the rationale behind any changes to the two versions of the notice.

23. A related issue is Ofcom's commitment to "take note of the information that services already include in their voluntary transparency reports and published reports required under other regulatory regimes. **This may be used to inform considerations of what information is feasible for services to collect and where it may be useful to require UK-specific versions of data that has already been published."** While we appreciate that very little transparency information has been published by services that is relevant to UK users, we are also concerned that this could be a route to further pushback from the providers - akin to requests under the FOI regime - where published information that looks broadly like the information requested is deemed to be sufficient and the granularity or specificity that Ofcom requires is overlooked, or that content does not get provided in the format requested.

24. We are happy to provide further information or discuss any of the above, if convenient.

October 2024
hello@onlinesafetyact.net

ANNEX: Evidence on limitations of current transparency reporting

Misleading claims about action on “anonymous” accounts

Following criticism for enabling racist abuse of England men’s team footballers during the 2021 Euros, in August 2021 Twitter put out this blog post. The post made numerous claims about the nature of the abuse and the perpetrators, and about actions which Twitter had taken. The apparent purpose of the post was to push back in general on the idea that the platforms could have done more, and in particular to push back on the idea that there was an issue with abuse from anonymous accounts which had been raised as an issue by several of the footballers themselves. Their post included a headline claim that “99% of the accounts suspended were not anonymous” i.e. 99% of the accounts which Twitter had identified as perpetrators weren’t anonymous.

These claims were widely reported, e.g. Time described them as Twitter “offering more transparency”. However, no evidence was offered to back up any of the claims. Clean Up The Internet wrote to Twitter requesting they provide more details to back up the claims, The UK’s Home Affairs Select Committee, and Labour MP Margaret Hodge also challenged them to back up these claims. It became clear that to arrive at the 99% figure, Twitter had used an extremely sporty definition of “anonymous” - an account called “Mickey Mouse”, linked to a disposable gmail address, would have been classified as “not anonymous”.

Full write up here:
https://www.cleanuptheinternet.org.uk/post/twitter-s-anonymity-claims-appear-to-rely-on-classifying-mickey-mouse-accounts-as-not-anonymous

Discrepancies in the TikTok researcher API

The TikTok researcher API delivered misleading results to researchers accessing via EU DSA - 2024: researchers realised that data accessed via the researcher API launched by TikTok in advance of DSA rules didn’t seem right, manually checked by what was publicly available or via scraping, and discovered discrepancies. More here:
https://www.techpolicy.press/-researcher-data-access-under-the-dsa-lessons-from-tiktoks-api-issues-during-the-2024-european-elections/

“Resolved” Meta complaint metrics not accounting for automated closure of reports

The Meta Oversight Board has identified instances where reports on posts were closed by automated systems as they did not receive a human review within 48 hours; these will have been recorded as “resolved” though no action was taken. See for example, here; here; and here. The automatic closure of reports/complaints is also an issue when applied to users who have appealed against a decision to have their post removed, too.

Meta's quarterly transparency reports stating a higher CSAM detection rate than reality

See Telegraph investigation here from 2021:
https://www.telegraph.co.uk/technology/2021/05/19/facebook-blamesglitch-huge-drop-child-abuse-image-takedowns/

Action against Snap in the US re fentanyl sales

Civil attorneys in the US are suing Snap Inc on behalf of families who lost children to deadly fentanyl tablets sold online. Snap tried to have the case dismissed referring to data in their transparency reports.

In a report from the Colorado Attorney-General, the following observations are made about transparency reporting about Fentanyl across a number of platforms: "even though some platforms have provided information about their efforts around drug activity, there remains a greater need for transparency and accountability in this regard. Platform responses generally have only provided highlights of their anti-drug actions, but lack any objective analysis of whether these approaches are effective and whether the efforts have successfully helped law enforcement and victim families take action against those continuing to use the platforms to distribute illicit substances online. Ultimately, independent external review is likely required to ensure that platforms are doing what is necessary to enforce their terms and community guidelines and are devoting enough resources to address the issue proactively."

Other concerns about Snap's transparency reporting and the robustness of their available metrics have been raised in the Judiciary Committee's inquiry (pp66-70), specifically regarding Snap's metrics on suicide and self-harm content. A further lawsuit against Snap has recently been filed by the New Mexico Attorney-GEneral alleging a number of failures by the company in protecting children from sextortion, sexual exploitation and other harms; the unredacted complaint details a number of areas where Snap's failed to act on evidence of harm on its platform, including

Lack of specificity on "actions" reported in metrics - large headline numbers obscure what action companies have actually taken and whether it has been effective

We provide here some analysis from an upcoming report from the Alliance to Counter Crime Online to the European Commission which looks at CSAM on Meta which we hope will be helpful for Ofcom in considering the metrics it will seek to require from the largest platforms:

> "While the DSA provides the first set of legally-enforceable transparency measures in the world, most major social media companies, including Meta, have issued voluntary transparency reports for years. However, critics, including ACCO, have long complained that Meta's transparency data is far from transparent, instead featuring carefully-chosen wording and statistics that make it seem like the

company is removing the vast majority of harmful content, and vague, general descriptions about what consequences actually occur if and when offensive content is actioned.

Since the DSA became enforceable, Meta began making an extraordinary admission on its transparency page: The company now specifically acknowledges that it doesn't know (or won't admit) how prevalent child endangerment violations are on Facebook or Instagram. "We will continue to expand prevalence measurement to more areas as we confirm accuracy and meaningful data."

Meta then goes on to explain that the company grades itself by what percentage of violating content its systems find and remove, compared to what its users report. This doesn't mean it's the only violating content on Meta platforms, just what is known about.

It's worthwhile to examine this methodology. Meta uses this same methodological approach to grade itself – no matter whether the content is illicit drugs, terror content or content that exploits children – measuring the percentage of violating content that its systems find and remove before users report it. ACCO has long argued that this approach

According to its own reporting, and assuming the company is being honest, Facebook's moderation systems were responsible for 94.3% of the 14.4 million actions the platform took against CSAM content in Q1 2024, the most recent quarter for which there was data when this report was prepared. Meanwhile, Instagram systems were responsible for 93.4% of the 2.7 million actions taken against CSAM on that platform.

These numbers may sound impressive, until one looks at them in reverse. The 5.7% and 6.6% failure rates mean that almost one million pieces of CSAM content were seen and reported by users before the company's systems found them, in just one quarter.

Again, that number may sound small when compared to the trillions of pieces of content shared daily on Meta platforms, however they are hardly trifling numbers from a law enforcement standard. To put those numbers in perspective, the U.S. Department of Justice seized the servers hosting Backpage.com for hosting child sex trafficking advertisements representing about two dozen plaintiffs. More recently, French authorities arrested Telegram CEO Pavel Durov after his platform refused to cooperate over requests to hand over data associated with accounts sharing CSAM and other illicit content."