Introduction

At Glitch, we are a UK charity dedicated to addressing online violence against women and girls, with particular focus on protecting Black women from digital harms.

Our important campaign to introduce violence against women and girls into the Online Safety Act has been core to our work over recent years and we are active participants in the current pre-consultation engagement led by Ofcom on the development of the VAWG guidelines. We have also actively engaged with Ofcom's Media Literacy strategy consultation and look forward to the development of this work moving forward.

Today, we welcome the opportunity to respond to Ofcom's consultation on draft transparency guidance. Transparency is essential for ensuring accountability in digital platforms, especially in the implementation of the Online Safety Act (OSA) and its subsequent impact on women and marginalised communities.

Below are our responses to the key areas where Ofcom is seeking input, along with additional recommendations for enhancing transparency and safety.

Glitch's key recommendations on transparency reporting

1. **Data should be requested on the basis of the prevalence and nature of harms, platforms' responses to those harms and platforms investment in work to prevent these harms**

2. **Transparency reporting should be required at set bi-annual reporting points, aligning with requirements of the EU Digital Services Act**

3. **By-and-for organisations should be resourced to continue engaging on this work, to ensure the voices of victim-survivors, particularly groups such as Black women who are disproportionately impacted by online harms, are working with Ofcom to iterate and improve this work.**

1. Ofcom's powers and providers' duties for transparency reporting, as well as Ofcom's approach to implementing the transparency regime

We recognise the critical role Ofcom plays in ensuring transparency under the OSA. However, for transparency reports to be effective, they must go beyond basic reporting and provide granular details that reflect the complex nature of online harms. Transparency reporting should be considered as part of a holistic approach to regulation of categorised services, therefore, in requiring transparency for categorised services, Ofcom should consider how such reports are made available to civil society, the public and academia.

In implementing the transparency regime, Ofcom should consider providing or requesting standardised data formats, so that they are publicly available and accessible to access and read.

Ofcom's approach for determining what information service providers should produce in their transparency reports

Data transparency on harms

- Characteristics of bad faith actors, including patterns in abusive behaviour.
- Granularity around policy violations, such as the specific types of violations (e.g., harassment, hate speech, misinformation).
- Granularity of the types of abuse reported (e.g., racist, misogynistic, xenophobic abuse) and whether it targets specific groups, such as Black women.
- Granularity of the types of hateful conduct flagged and reported, distinguishing between user-flagged and algorithmically flagged content.
- Granularity of the types of accounts taken down, including information on the nature of those accounts (e.g., bots, fake accounts, or individual users).
- Granularity of how accounts were taken down, such as whether the removal was due to algorithmic detection, third-party reporting (trusted flaggers), or personal reporting. This would help assess the effectiveness of bystander reporting and automated systems.
- Reports on spikes in harmful content during significant political or social events (e.g., elections, protests, cultural moments).

Data transparency on response to harms

- Timeframes for account removals: how long it takes accounts to be taken down from the moment they are reported
- The number and proportion of reported content by users that does not meet the threshold for moderation

- Country comparisons to understand differences in safety measures and abuses across regions.
- Data on the role of recommender algorithms in amplifying harmful content and how this is tracked.
- Content moderation- granularity for metrics on accuracy metrics on automated enforcement tools
    - Comparison on this accuracy of automation vs human review of illegal and harmful content
- Design features- Platforms should report how Safety by Design elements, such as block and mute features, help mitigate harmful content, broken down by geography, language, gender, sex and race, paying attention to marginalised groups, specifically women of colour and Black women.

Data transparency on prevention of harms

- Which and how many internal policies are currently deployed by companies in their trust and safety work. For example as a previous member of Twitter's Trust and Safety Council we helped to develop their "Hateful Conduct Policy" but this has since been revoked in practical content moderation, so it would be helpful to know which policies do exist internally.
- Safety feature usage statistics, showing how many people are using features like block, mute, and filter functions.
- Details on the workforce involved in trust and safety, including the number of people, their location, languages and geographies prioritised, and information on whether these teams are outsourced.
- Overview of current priorities of Trust and Safety teams internally.
- The ways that users make use of platform affordances providing control and/or adjustment on recommender algorithms, including qualitative summary on how *effective* algorithms are at reducing harmful content
- Data on how platforms create and promote digital literacy opportunities and resources for users, including for example reporting tools are accessible and effective for all users, evidenced via usage data.

Online gender-based violence transparency data

- A detailed breakdown of incidents (e.g. kinds of abuse, account removals) relating to online gender-based violence (OGBV).

- Platforms should report on the following as it relates to online gender-based violence (OGBV): the type of abuse (e.g., misogynistic slurs, harassment, stalking, image-based abuse); the frequency of incidents, and the specific demographics of those targeted. These kinds of abuse would fall under the OSA's definition of 'illegal content'.
    - Particular attention should be given to incidents of abuse where individuals face multiple forms of discrimination (e.g., racism and sexism) simultaneously.
    - Reports should highlight how often racialised women are specifically targeted with racist and sexist attacks.
    - With regards to OGBV specific reports should indicate quantitatively, and qualitatively, the instances in which Black women are subject to various forms of digital misogynoir.
    - Reports on OGBV in general should track how these incidents are flagged, reported, and addressed by platforms. This data should also cover the *outcomes* of reports, (i.e., whether content was removed, accounts were banned, or further actions were taken) and the timeframes involved in resolving such cases.
    - Lastly, platforms should disclose the measures they have in place to address forms of racialised OGBV, intersectional abuse and their concrete activities towards protecting women, with particular attention to Black women from targeted online violence.

2.1 Timings for Implementation: Transparency Reporting

We believe it is important for transparency reporting to be structured in a way that ensures data is consistently collected and reported in a timely manner. To this end, we recommend the following:

- Fixed Date for Transparency Reports: Instead of issuing annual transparency notices at different times, we propose setting fixed dates for bi-annual submission of transparency reports—in line with best practice under the Digital Services Act (EU). This would ensure that reports from all service providers are aligned for comparison.
- Tailored Timelines Based on Provider Size: Based on provider feedback, a timeline can be agreed upon for data submission, depending on the size and capacity of the provider. For instance, larger providers may submit within two months, while smaller ones may be given more flexibility (up to six months).
- Continuous Data Collection: An initial set of basic requirements can be created for transparency reports, with more specific requirements added later (with three months'

notice). This would encourage continuous data collection throughout the year, ensuring comprehensive and up-to-date reporting.

3.38d Engaging with Stakeholders and Experts to Iterate and Improve the Transparency Regime

We strongly believe that user voices and perspectives should be considered primary stakeholders in the transparency process. This is especially true for victim/survivors of online harm, who must be part of engagement processes via specialist for-and-by organisations who must be financially resourced to engage in this work. Transparency reporting should reflect the lived experiences of those who have suffered from online abuse, ensuring that the insights gained from these users guide future improvements to platform safety.

3.42 Measuring Digital Safety

We need further clarification on the weightings and thresholds for assessing "Impact," "Risk," and "Process" in Ofcom's digital safety evaluation. Specifically, how will these factors be measured and applied in relation to service providers? Understanding the framework for assessing risk and impact will be critical for holding providers accountable.

4.3d Engagement with Providers

While the proposed engagement activities for service providers are a positive step, there is a lack of clarity on how these activities will be carried out on the ground. It is essential that providers understand their duties and Ofcom's expectations, particularly in how they engage with marginalised communities and respond to their specific needs. Ofcom should consider providing more detailed guidance on how this engagement will take place, including how feedback from civil society groups and impacted communities will be integrated into policy and enforcement decisions.

Conclusion

We commend Ofcom's commitment to transparency and accountability in the digital space. However, we believe that the inclusion of clearer data requests, pre-set bi-annual reporting, and a stronger focus on user voices, particularly those of marginalised communities, is essential to creating a safer online environment. We look forward to continued engagement and collaboration with Ofcom to ensure that the Online Safety Act is implemented in a way that truly protects vulnerable users from online harm.