Introduction

1. ISD welcomes the opportunity to respond to this important call for evidence on Ofcom's draft guidance on transparency reporting duties for categorised services. We also recognise the challenges the complexity of the Online Safety Act (OSA) brings, especially in the context of the collective desire from the Government, Ofcom, and other key stakeholders to quickly and effectively implement and enforce the legislation and improve online safety in the UK.

2. Given the lack of mandatory researcher data access provisions in the OSA, the transparency reporting requirements are an even more crucial tool to enable broader public understanding of the online environment in the UK and the various risks to safety that online platforms can present. In the absence of mandatory data access, both services' and Ofcom's transparency reports will play a vital role in the work of wider non-industry stakeholders – whether in civil society or academia – to scrutinise both online services and the work of Ofcom as the regulator. This work is crucial to informing the Government and other political stakeholders, media reporting of online harms, and the wider public.

3. As stated in our previous consultation responses to Ofcom's illegal harms and categorisation proposals, in our view, many services' existing transparency reports often rely on self-selected metrics and measures of success that do not present an objective assessment. A notable lack of transparency, evidenced by self-defined or self-serving metrics, intentional gaps in transparency disclosures, and evasive behaviour before Congressional or Parliamentary committees, underscores companies' reluctance to be subject to independent scrutiny and oversight.

4. There are also numerous examples of companies restricting access to data for independent researchers, including Twitter/X revoking free access to its API and taking legal action against the Centre for Countering Digital Hate (CCDH), and Meta withdrawing support for CrowdTangle and shutting down New York University's Cybersecurity for Democracy project's access to Facebook's Ad Library. In light of these challenges, and to meet its stated online safety objectives, Ofcom must ensure that its transparency reporting guidance does not simply consolidate existing approaches from industry, particularly for the highest risk services regardless of their size.

Categorisation

5. As previously noted in our response to Ofcom's consultation on the categorisation of services, we share the concerns outlined by the Online Safety Act Network here that Ofcom is not making use of the additional flexibility provided in the final Act that allows either.size or.functionality to be considered when assessing the levels of risk a service presents, and therefore the types of duties they should be subject to. Ultimately the intent of the Act is to effectively mitigate risks online, and we are concerned these proposals will leave important loopholes for certain small but high-risk services that will not be categorised.

6. In the context of the types of harms that ISD focuses on, including online terrorism, (violent) extremism, hate speech and targeted harassment, cross-platform dynamics and the

interconnected nature of the online ecosystem of platforms and services mean that harmful content or activity is often initially disseminated or coordinated on smaller platforms before migrating to larger platforms.

7. In instances where a service could present high levels of risk, offer relevant types of functionalities, but fall short of the proposed user number thresholds, we would be concerned that they would escape categorisation under Category 1 or 2b. This would mean they would be exempt from important additional duties, such as transparency reporting, which could enhance user safety, both on and off-platform.

8. We would therefore argue that this would be a missed opportunity to better scrutinise the safety measures in place on smaller platforms that play a key role in online extremism and hate, and address these harms at-source, rather than waiting for their impact to be magnified on larger platforms.

Proportionality

9. While we recognise that transparency notices should be proportionate and specific to each service, we would also stress that this should not be a major obstacle to requesting the detailed and granular information required to fully understand the nature, extent and severity of online harms of different services, and the effectiveness (or lack of) of their safety measures to address them, particularly for some of the largest and most profitable companies in the sector.

10. We would also argue that any assessment of proportionality must significantly evolve over time, as once a service has invested the initial resources necessary and is set up to produce the required transparency information, then it should be easier and more cost-effective for this information to be produced on an ongoing basis. As a result, it could then be proportionate to request significant additional information over time that may not have been deemed proportionate initially.

11. We believe this is particularly important in the context of Ofcom's proposed approach to 'thematic' reporting to ensure that as Ofcom's focus and priorities change, that there is not a risk of certain information no longer being required and services therefore backsliding on their efforts to address certain online harms.

Ofcom's proposed approach to draft transparency notices, industry engagement, and engagement with non-industry stakeholders

12. We recognise the intent behind Ofcom's proposed approach to sharing draft transparency notices with services is to ensure that they have a clear understanding of the information required and are able to clarify this where necessary prior to the formal notice being issued.

13. However, we are concerned that this additional step, which is not explicitly required under the OSA, will provide an opportunity for industry to make representations in private that could lead to the watering down of the requests on the basis of feasibility, relevance or proportionality. We are concerned that these representations may be difficult for Ofcom to

accurately assess and adjudicate, and would not provide an opportunity for other key stakeholders to provide input in the same way.

14. We would therefore recommend that Ofcom conducts this process transparently, including by publishing any changes made between the draft and final formal notices, and seeks more formal input from non-industry stakeholders, such as experts in civil society and/or academia.

15. The consultation documents also mention several potential approaches to engaging with wider online safety stakeholders, such as civil society and academia. Based on our experiences in an EU Digital Services Act (DSA) context to-date, we would recommend that Ofcom ensures that these engagements are regular, well planned and structured, conducted transparently and openly as far as possible, and allow stakeholders to provide input on the agendas and topics discussed (e.g. for events, meetings etc).

Formats and accessibility of transparency reporting

16. In order for the OSA transparency reports to make a meaningful impact on Ofcom's overall online safety objectives, they will need to be as accessible as possible. Currently, many services publish their transparency reporting in formats (e.g. PDF) that do not allow for the easy extraction and analysis of the included and underlying data. We would strongly recommend that Ofcom requires services' transparency reports to be published in a way that enables non-industry stakeholders to further make use of the data included, for example by publishing spreadsheets (e.g. CSV files) that can be used for additional analysis and cross-platform comparisons.

17. To ensure transparency reporting is accessible to the broadest possible audience, including members of the public and civil society organisations that do not directly conduct research on online services, we would also recommend that the largest and most popular services (particularly with children) are encouraged to produce public-facing, accessible and interactive visualisations or dashboards alongside static reports. We believe that this would allow for greater engagement with the content of the reports, and make it easier for non-experts to understand their content and potential implications (e.g. parents).

18. Similarly, we would also recommend that Ofcom develops an online transparency centre, portal or repository that links to all services' reports in one place (as they are not always easy to find on certain services, which reduces the likelihood of non-experts seeking them out). This resource could also host the underlying data in a more dynamic way that allows for further independent analysis (see for example, the dashboards provided as part of the EU's DSA Transparency Database that contains the 'statement of reasons' services are required to provide when moderating content). This could allow parents for example to be able to quickly and easily compare the prevalence of certain types of harmful content and how effectively it is addressed across popular platforms. The accessibility of this information will be crucial to develop and enhance public understanding, which is a key objective of Ofcom's proposals.

Methodologies, metrics and indicators

19. Given the breadth of the areas covered under Parts 1 and 2 of Schedule 8 in the OSA, and without the register of categorised services, it is difficult to provide a comprehensive and exhaustive set of key metrics or indicators across harms and platforms at this stage. This is compounded by the historical lack of transparency from services, which means that it is not possible to fully understand the types of data that services already collect and possess internally, and therefore make an informed assessment of the types of information and/or data that would be proportionate to request. Often, leaked or subpoenaed documents and the testimony of whistleblowers has demonstrated that services have access to far more comprehensive and detailed internal information (including specific research on the impacts of their products) than they have revealed publicly in their transparency reporting or user-facing explainers, blog posts etc.

20. However, overall we would strongly encourage Ofcom to request that data is provided in a format that is broken down and cross-referenced as far as possible to ensure a sufficient level of nuance, and that the way data is presented does not obscure underlying relationships, trends or anomalies. This should include the following:

    A. **Types and subcategories of harm**: While the OSA covers a wide range of different illegal offenses and harms to children, as far as possible data should be subcategorised. For example, to fully understand the nature and extent of illegal hate speech on a service, it would be important to understand the specific groups targeted by such content.

    B. **Timescales**: If data is available to services hourly or daily, then it should not only.be provided in a monthly or quarterly format in their reporting, otherwise there is a risk that this prevents closer scrutiny of services systems and processes (e.g. time taken to moderate content at different times of day, over weekends etc), or the effectiveness of their safety measures during particular moments, events or crises that are likely to be relevant under the Act (e.g. a terrorist attack, civil unrest/riots, elections, sporting events etc.).

    C. **Demographics**: To fully understand the extent to which users encounter or are exposed to illegal or harmful content, data must be provided in a way that allows for closer analysis of the types of users being exposed to different forms of illegal or harmful content – e.g. on the basis of age, gender etc. This data should also allow for analysis on the distribution of exposure, i.e. are a small number or specific group of users being consistently exposed, or are a larger number being exposed more incidentally.

    D. **Engagement**: Although 'views' or 'impressions' can be useful metrics to understand exposure risks, depending on how they are defined by a service they can be misleading in terms of the extent to users actually engage or interact with content. It would therefore be useful for services to provide more detailed engagement data (e.g. likes, comments, shares, view time etc.) to enable a deeper understanding of the potential impact of illegal or harmful content.

    E. **Languages**: Data on the prevalence, dissemination or engagement with content and its moderation should be language-specific to allow for more detailed analysis of the experience of users that speak languages other than English in the UK.

F. **Features and functionalities**: As far as possible, data should be provided that is sub-divided according to the specific features and functionalities of each service. For example, on a video sharing platform such as YouTube, it would be important to be able to understand prevalence, exposure and content moderation statistics across different aspects of the platform, e.g. standard videos, YouTube Shorts, livestreamed video, comments etc.

G. **Recommendation systems**: This approach should also extend to algorithmic systems and recommendations – i.e. detailed metrics that provide a deeper understanding of how users arrive at certain illegal or harmful content from different types of recommendations (i.e. from the home page, 'watch next', autoplay, search, playlists etc.) or external sources (i.e. URLs elsewhere online). Where possible, this should also include information on the recommendation of accounts, channels, pages etc. that are associated with illegal or harmful content or activity, which is crucial to understanding certain forms of online harm.

H. **User safety and reporting tools**: Transparency reporting should include detailed information on the usage of safety and reporting tools by different groups of users – i.e. broken down by age, gender etc. – to better understand any accessibility challenges and the extent to which they are effective in mitigating different types of risks. Data should also be included that illustrates the types of content or accounts that users are proactively reporting, how accurately they are doing so, and the resulting actions taken by services to these reports. This type of data can be used to understand how well users understand services terms and conditions and policies (i.e. Community Standards or Guidelines) and inform digital and media literacy efforts.

I. **Safety measures or mitigations**: Deeper levels of granularity that go beyond what is already provided by services will be crucial to understand the effectiveness and impact of a wide variety of different content moderation and other safety measures employed by services. For example, it is not sufficient to services to provide data on illegal or harmful content that has been 'Actioned', without a more detailed breakdown of what action has been taken (i.e. content or account removal, geo-blocking, labelling, de-amplification, account warning etc.). Services should also provide clear information on the extent to which content moderation is conducted through automated means, and the relative accuracy and effectiveness of such approaches in comparison to human review.

21. Overall, it will also be vital that services are required to provide UK-specific information that does not obscure or conflate the extent to which services are dedicating resources to the UK context. For example, providing the number of English-speaking moderation staff worldwide is not a helpful indicator of the extent to which they are addressing content impacting UK users.

22. Services should also be required to include detailed information on their methodological approach to producing their transparency reporting. This should include detailed definitions for each metric describing how they are calculated, as well as any potential limitations or caveats to the data provided. Overall, it is vital that Ofcom drives a cultural shift in the way

services report transparency data towards a more open and honest approach that does not try to obscure weaknesses or failings in their approaches to online safety.