Question	Your response
We welcome input from industry on the areas listed below. We encourage stakeholders to respond with feedback so that we can ensure that the guidance helps providers and other stakeholders understand:	
A) Ofcom's powers and providers' duties for transparency reporting, as well as Ofcom's approach to implementing the transparency regime. B) Ofcom's approach for determining what information service providers should produce in their transparency reports. C) Ofcom's plans to engage with providers prior to issuing transparency notices, and on what matters, and whether the proposed engagement plan will be sufficient for helping services to comply with their duties. D) Ofcom's plans to use the information in providers' transparency reports in Ofcom's own transparency reports.	Confidential? N Big Brother Watch welcomes Ofcom's measures to improve transparency. Tech companies operate with almost unchecked power and transparency reports are among the most effective tools that we have at our disposal to encourage accountability and uphold the rights of users. Without service providers disclosing relevant information, stakeholders are unable to examine the effects of platforms' interventions and their implications for privacy and free expression. Ofcom should be intentional about what information it requires from tech companies in order to make its attempts at greater transparency meaningful. Ofcom should require that platforms report on their human rights and due process considerations; the range of actions the service provider may take against user content and accounts due to violations of their rules and policies; complaints and appeals mechanisms; state involvement in flagging and content moderation, and the accuracy of their systems and external auditing, as per the Santa Clara Principles 2.0 on how best to obtain meaningful transparency and accountability from platforms.[1] These principles were developed in 2018 by a group of human rights organisations and academic experts to establish meaningful accountability and transparency around service providers' content moderation practices. Since 2018, twelve major tech companies, including Meta, Google and Apple, have endorsed the principles, which were further expanded in the second

[1] https://santaclaraprinciples.org/

	iteration. Despite this endorsement, the principles have not been adequately reflected in these companies' practices, and this is where the Ofcom requirements can make a difference. In line with these principles, human rights and due process should be integrated at all stages of the content moderation process and service providers should publish information about how they achieve this aim.[2] We understand that Parts 1 and 2 of Schedule 8 of the OSA grant Ofcom a broad legal remit to request information from service providers and that the Regulator will decide what to request on the basis of its relevance, appropriateness and proportionality. The proposals we make for disclosure are relevant, appropriate and proportionate, as section 22 of the OSA contains cross-cutting duties about privacy and free expression. We have organised our response around three areas of disclosure recommended by the Brookings Institute:[3] due process protections, material subject to moderation, and algorithms. **1. Due Process** Whilst we welcome any measures that encourage greater transparency and accountability in relation to the content moderation practices employed by tech companies, it is our view that annual reporting does not go far enough. Section 21 of the OSA sets out duties on service providers in relation to the operation of their complaints procedures. Whilst a duty on platforms to integrate complaints processes, as required by s.21 of the Online Safety Act,

[2] Santa Clara Principles 2.0, Principle 1

[3] https://www.brookings.edu/articles/how-online-platform-transparency-can-improve-content-moderation-and-algorithmic-performance/

	is a welcome step when it comes to protecting freedom of expression online, the reality is that many platforms already offer variations of this function, which in many cases lacks transparency or rigour. The legislation does not include any provisions to improve or set minimum standards for these complaints processes. Further, this measure will make little difference if the bar for what is considered acceptable online is considerably lowered. Ofcom should provide more specific guidance about what platforms' appeal processes should look like, including a requirement for human review and a detailed explanation of the outcome. The Regulator should also require providers to be more forthcoming to users on a case-by-case basis about exactly what action has been taken in relation to each piece of uploaded content that is subject to content moderation. This measure would be in line with Article 17 of the EU's Digital Services Act (DSA), which requires platforms to "provide users with a clear and specific statement of reasons" as to why a user's content was moderated, in cases where it breached the platform's terms of service or was illegal content. The statement of reasons should include an explanation of which content rule the offending material breaches, how the content will be dealt with (i.e., removed, down-ranked or delayed), whether it was flagged using AI detection or by a user, and whether the content moderation decision was taken by an automated system or human review. As many of the designated service providers will, no doubt, operate internationally, Ofcom making similar requirements of platforms should not create additional burdens. In order to provide civil society with a clear picture of the type of material being removed and down-ranked from social media, there should also be public disclosure of content moderation data. We recommend that Ofcom should require service providers to collate and submit their statement of

	reasons, in line with Article 24(5) of the DSA, which then forms a publicly accessible database for research and analysis.[4] Any personal information should be redacted to protect the privacy of those affected.[5] **2. Content Moderation** Disclosures on the number of content removal actions taken by services are insufficient. We need more granular data about the types of content that have been censored so the accuracy of content moderation and the true extent of restrictions on free speech can be accurately assessed. Ofcom should mandate service providers to disclose the number of pieces and type of content on which they take action, the type of action taken and how the content was detected. The Santa Clara Principles 2.0 state that "Companies should report information that reflects the whole suite of actions the company may take against user content and accounts due to violations of company rules and policies, so that users and researchers understand and trust the systems in place." In line with these Principles, service providers should disclose the number of successful and unsuccessful appeals that resulted in pieces of content or accounts being reinstated; that were initially flagged by automated detection; and that were reinstated without appeal after being erroneously actioned.[6] As aforementioned, accumulating the statement of reasons in a database will also provide transparency over the types of content subject to moderation and the reasons for interventions. Without such disclosure, our awareness of the type of content being restricted will be dependent on individual disclosures and the outcome of complaints. This clearly is not enough to

[4] https://www.law.kuleuven.be/citip/blog/the-digital-services-act-towards-more-transparency-for-content-moderation/

[5] https://www.law.kuleuven.be/citip/blog/the-digital-services-act-towards-more-transparency-for-content-moderation/

[6] Santa Clara Principles 2.0, Operational Principles.

	provide a picture of the state of free speech online. In other consultation documents, Ofcom has lauded the role that "trusted flaggers" can play in content moderation processes. However, Big Brother Watch's research into the UK government's counter-disinformation units (operating out of various government departments) uncovered a worryingly close relationship between civil servants and social media companies, with companies being informally pressured to remove content that was lawful, raising wider concerns about the extent to which these relationships between state bodies and social media platforms are both transparent and rights-respecting.[7] In its efforts to improve accountability and transparency, Ofcom should impose duties on service providers to disclose information about this relationship. As the counter-disinformation units show, the informal nature of civil servants' requests mean that Ofcom should require service providers to go beyond just providing information about formal legal orders from state authorities and include information about the number and nature of content flags from all representatives of the state. The Santa Clara Principles 2.0 state that users should know when a state actor has requested or participated in any actioning on their content or account and whether the intervention was required by law. Additionally, users should be able to access "details of any formal or informal working relationships and/or agreements" between the service provider and state actors in relation to flagging content, accounts and any other actions taken.[8] This aligns with the recommendation we made in our Ministry of Truth report that any government correspondence with an online intermediary

[7] Ministry of Truth – Big Brother Watch, January 2023: https://bigbrotherwatch.org.uk/wp-content/uploads/2023/01/Ministry-of-Truth-Big-Brother-Watch-290123.pdf

[8] Principle 4 of the Santa Clara Principles 2.0

	regarding specific pieces of lawful content on their site should be made public.[9] **3. Algorithms** In order for civil society to be able to understand how the infrastructure of these platforms affects the service provided and how individuals' legal rights are engaged, we need to be able to analyse the algorithms they employ, including content moderation and recommender systems. Ofcom should require service providers to explain how content decisions are made, particularly whether they were made by humans or automated systems. Where automated systems are used, the Santa Clara Principles 2.0 recommend that service providers should disclose when, how and on what types of content they are deployed; the accuracy rates including differences between languages and categories of content; the criteria for decision-making; and the number of successful and unsuccessful appeals where the content was initially automatically detected. The Think Tank, New America recommends that such disclosure should extend to the types of information that datasets contain, including how regionally, linguistically, and demographically diverse the data are, what outputs the models generate, and the accuracy rates of human and automated decisions.[10] Principle 5 of the Santa Clara Principles 2.0 emphasises that service providers should publish information regarding the accuracy of their systems and submit their process and algorithmic systems to periodic external auditing. We would welcome these measures to allow for decision-makers, researchers, civil society

[9] https://bigbrotherwatch.org.uk/wp-content/uploads/2023/01/Ministry-of-Truth-Big-Brother-Watch-290123.pdf, p93

[10] https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/promoting-fairness-accountability-and-transparency-around-automated-content-moderation-practices/

	and users to independently assess and scrutinise how speech is being moderated online. We additionally support the measures which allow for "comparisons between services" to be made. We support the provisions about bespoke requirements, as adopting an overly standardised approach could result in the omission of relevant information. <u>Section C</u> We support the commitment to "dedicated engagement with civil society groups." We also support the international approach as it is helpful to understand whether the measures are overly restricting access to information for UK users as compared to around the world. <u>Section D</u> Ofcom and policy-makers should be aware that Ofcom's own transparency reports will be based on the information supplied by service providers and will therefore have limitations.
Are there any aspects in the draft guidance where it would be helpful for additional detail or clarity to be provided?	Confidential? - N
Are the suggested engagement activities set out in the draft guidance sufficient for providers to understand their duties and Ofcom's expectations?	Confidential? - N

Question	Your response
We are also seeking input that will help us understand if there are other matters that Ofcom should consider in our approach to determining the notices, beyond those that we set out in the guidance. The questions below seek input about any additional factors Ofcom should take into account in various stages of the process, including: to inform the content of transparency notices; in determining the format of providers' transparency reports; and how the capacity of a provider can be best determined and evidenced.	
Are there any other factors that Ofcom might consider in our approach to determining the contents of notices that are not set out in the draft guidance?	Confidential? - N

Is there anything that Ofcom should have regard to (other than the factors discussed in the draft guidance) that may be relevant to the production of provider transparency reports? This might include factors that we should consider when deciding how much time to give providers to publish their transparency reports.	Confidential? - N
What are the anticipated dependencies for producing transparency reports including in relation to any internal administrative processes and governance which may affect the timelines for producing reports? What information would be most useful for Ofcom to consider when assessing a provider's "capacity", by which we mean, the financial resources of the provider, and the level of technical expertise which is available to the service provider given its size and financial resources?	Confidential? - N
Are there any matters within Schedule 8, Parts 1 and 2 of Act that may pose risks relating to confidentiality or commercial sensitivity as regards service providers, services or service users if published?	

Question	Your response
Finally, we are also seeking input into any matter that may be helpful for ensuring Ofcom's transparency reports are useful and accessible.	
Beyond the requirements of the Act, are there any forms of insight that it would be useful for Ofcom to include in our own transparency reports? Why would that information be useful and how could you or a third party use it?	Confidential? - N
Do you have any comment on the most useful format(s) of services' transparency reports or Ofcom's transparency reports? How can Ofcom ensure that its own transparency reports are accessible? Provide specific evidence, if possible, of which formats are particularly effective for which audiences.	Confidential? - N

Question	Your response

Please provide any other comments you may have.	
General comments	Confidential? - N

Please complete this form in full and return to OS-Transparency@Ofcom.org.uk