Consultation response form

Please complete this form in full and return to protectingchildren@ofcom.org.uk.

Consultation title	Consultation: Protecting children from harms online
Full name	Jasleen Chaggar
Contact phone number	020 8075 8480
Representing (delete as appropriate)	~~Self /~~ Organisation
Organisation name	Big Brother Watch
Email address	jasleen.chaggar@bigbrotherwatch.org .uk

Confidentiality

We ask for your contact details along with your response so that we can engage with you on this consultation. For further information about how Ofcom handles your personal information and your corresponding rights, see Ofcom's General Privacy Statement.

Your details: We will keep your contact number and email address confidential. Is there anything else you want to keep confidential? Delete as appropriate.	Nothing / ~~Your name / Organisation name / Whole response / Part of the response (you will need to indicate which question responses are confidential)~~
Your response: Please indicate how much of your response you want to keep confidential. Delete as appropriate.	None / ~~Whole response / Part of the response (you will need to indicate below which question responses are confidential)~~
For confidential responses, can Ofcom publish a reference to the contents of your response?	Yes / ~~No~~

Your response

Question	Your response

Volume 2: Identifying the services children are using Children's Access Assessments (Section 4).	
Do you agree with our proposals in relation to children's access assessments, in particular the aspects below. Please provide evidence to support your view. 1. Our proposal that service providers should only conclude that children are not normally able to access a service where they are using highly effective age assurance? 2. Our proposed approach to the child user condition, including our proposed interpretation of "significant number of users who are children" and the factors that service providers consider in assessing whether the child user condition is met? 3. Our proposed approach to the process for children's access assessments?	Big Brother Watch is committed to defending the privacy and free speech rights of all internet users. We believe that more can and should be done to protect children online in a way that respects their rights. As an independent regulator, we recognise the legal limitations that bind Ofcom's approach when planning the implementation of the Online Safety Act. However, we have taken the opportunity to respond to this consultation in a way which highlights those areas of the Online Safety regime which engage the rights to free speech and privacy, especially where Ofcom has used its own discretion in implementing duties set out in the legislation. As a public body, Ofcom is also bound by its obligations set out in the Human Rights Act 1998 and it is vital that the rights to free speech and privacy, protected by articles 10 and 8 of the Act respectively are not compromised by measures intended to keep children safe online. We opposed the inclusion of provisions in the Act that stipulate that a provider is only entitled to conclude that children cannot access a service if age verification or age estimation is used (s35(2)).As a result, services must decide between forcing users to employ invasive age verification or estimation technology, and implementing extensive content moderation tools to censor their content to make it appropriate for users under the age of 18. We firmly oppose this approach and believe that it will

	engage individuals' rights to both free expression and privacy. Age verification controls can be easily circumvented and should not be seen as a silver bullet solution - parental controls,user controls and age rating are other recognised, reliable methods to protect children from inappropriate content online.[1] Our broader objections to highly effective age assurance (HEAA) are discussed below in response to question 31, however we are concerned that allowing services who implement HEAA to bypass the need to complete a risk assessment and other child safety duties will incentivise providers to use HEAA to avoid the administrative burdens associated with not using it. The guidance acknowledges that "most part 3 services are likely to meet the child user condition" (Vol. 2, para 4.44) and that "this will result in small, low-risk services incurring costs of conducting a children's risk assessment" (Vol. 2, para 4.67). As a result, most services – to avoid the intricacies, costs and pitfalls of completing and implementing risk assessments and other obligations – will instead opt for HEAA. This could lead to a disproportionate outcome whereby services that are unlikely to host content that is harmful to children nonetheless employ HEAA, which in turn effectively creates an ID requirement for the internet and makes measures AA1-AA6 redundant. We disagree with the definition of a significant number of children, which Ofcom suggests

1https://blogs.lse.ac.uk/parenting4digitalfuture/2021/11/17/age-assurance/

	could include "even a relatively small absolute number or proportion of children could be significant in terms of the risk of harm to children" (Vol. 2, para 4.23). This goes against the ordinary meaning of 'significant number' (s35(3) of the Act) and demonstrates an instance of Ofcom using its discretion over the implementation of the Act which would make measures likely to erode individuals' privacy online more likely.
Volume 3: The causes and impacts of online harm to children **Draft Children's Register of Risk (Section 7)**	
Proposed approach: 4. Do you have any views on Ofcom's assessment of the causes and impacts of online harms? Please provide evidence to support your answer. a. Do you think we have missed anything important in our analysis? 5. Do you have any views about our interpretation of the links between risk factors and different kinds of content harmful to children? Please provide evidence to support your answer. 6. Do you have any views on the age groups we recommended for assessing risk by age? Please provide evidence to support your answer. 7. Do you have any views on our interpretation of non-designated content or our approach to identifying non-designated content? Please provide evidence to support your answer. **Evidence gathering for future work:** 8. Do you have any evidence relating	We are cognisant of the fact that as an independent regulator, Ofcom is bound by the duties conferred to it by the Online Safety Act. However, we have taken the opportunity to highlight a number of assumptions Ofcom have made in its diagnostics of "online harm" and the impact of different digital infrastructures which then attempt to justify further action which may have a bearing on free expression and privacy online. For example, Ofcom's approach to online abuse is one which poses that anonymity is a causative factor in the online harms experienced by young people. The proposals suggest that the "far greater potential for anonymity online may enable users to trivialise the consequences of their actions and break social norms of respect and decency that they may adhere to in-person interactions" (Vol. 3, para 7.4). We would emphasise the important role that anonymity holds in allowing individuals to navigate the online world without identifying themselves, in particular related to mental health, sexuality and abuse around the world. Online anonymity is equally valuable to LGBTQ people who may wish to navigate the internet anonymously to explore their identity, as well as survivors of domestic abuse who might seek support without revealing their identity. In 2015, the former United Nations special rapporteur on freedom of expression reported "Encryption and anonymity

to kinds of content that increase the risk of harm from Primary Priority, Priority or Non-designated Content, when viewed in combination (to be considered as part of cumulative harm)? 9. Have you identified risks to children from GenAI content or applications on U2U or Search services? a) Please Provide any information about any risks identified 10. Do you have any specific evidence relevant to our assessment of body image content and depressive content as kinds of non-designated content? Specifically, we are interested in: a) (i) specific examples of body image or depressive content linked to significant harms to children, b. (ii) evidence distinguishing body image or depressive content from existing categories of priority or primary priority content. 11. Do you propose any other category of content that could meet the definition of NDC under the Act at this stage? Please provide evidence to support your answer.	provide individuals and groups with a zone of privacy online to hold opinions and exercise freedom of expression without arbitrary and unlawful interference or attacks".[2] There is little evidence to suggest that anonymity itself makes online discourse more febrile. Further, online anonymity or pseudo-anonymity is not a barrier to tracking down and prosecuting those who commit criminal activity on the internet. Police reporting shows that in 2017/18, 96% of attempts by public authorities to identify the anonymous user of a social media account, email address or telephone, resulted in successful identification of the suspect of their investigation. Ofcom identifies a wide, all-encompassing range of functions (including livestreaming, group messaging, encrypted messaging, commenting on content, posting content, reposting or forwarding content, and reacting to content) which, it suggests, can increase the risk of children being exposed to harmful content. This seems to be a catch-all approach with every type of communication between children being labelled as potentially dangerous. This characterisation runs the risk of pressuring providers to over-censor and excessively surveill online content. Arguably, the particularly damaging feature in relation to children's safety is the algorithmic predictive content loop. This should be the target of restrictions, rather than free speech. We agree that recommender systems and advertising-based business models are a risk factor can contribute to harm, since they remove control away from children over what they see. In our view, children and their parents – not recommender systems or content moderation software – are best placed to decide what they should see on the services they interact with. We are concerned by several assumptions made by Ofcom that are not necessarily bourne out by evidence. Ofcom explains that it

2Special Rapporteur on freedom of opinion and expression, Report on encryption, anonymity, and the human rights framework, 22 May 2015, p7: https://documents.un.org/doc/undoc/gen/g15/095/85/pdf/g1509585.pdf?token=KljzySMTzNIuWumoCH&fe=true

	has "made reasonable inferences about the risks that may arise" on search services in instances where it does not have "specific evidence about that service type," (Vol. 3, para 7.38(d)). We alarmed by these unsubstantiated assumptions being used to guide policy. Given the importance and implications of the measures that Ofcom is proposing, we do not agree that they should be adopted on the basis of inference. We have previously raised concerns about the way in which social media platforms define the term "hate speech" and this is reflected in assertions Ofcom make about "online abuse and hate content". Ofcom posits that hate speech online can encourage individuals to "adopt hateful attitudes" and is associated with the "normalisation of discriminatory attitudes" (Vol 3, paras 7.4 and 7.4.26). However, in order to have meaning, hate speech is a term that requires precise definition. The UK has expansive laws governing speech-related offences that can be used to prosecute violent, hateful and harmful forms of speech and behaviour online. Yet platforms' definitions of "hate speech" do not align with UK domestic law and can result in the censorship of online speech which is entirely lawful and reasonable. Our research has shown that in debates about hate speech or online abuse, light-hearted and sarcastic comments can be inaccurately characterised as causing harm. For instance, the comic, Marcia Belsky, was banned from Facebook for 30 days in Autumn 2017 after commenting "men are scum" on another user's post discussing her experience of misogyny.[3] Ofcom's assessment of the causes and impacts of self-harm content also adopts a flawed approach. There is no question that content encouraging or assisting suicide should be removed, and indeed encouraging and assisting suicide is an offence, which the CPS is clear applies equally to content online. However, Ofcom's proposals seek to "shield" internet users from all material relating to self-

3 Marcia Belsky, Twitter, https://twitter.com/MarciaBelsky/status/921082758574854146; Big Brother Watch, 'The State of Free Speech Online,' September 2021, https://bigbrotherwatch.org.uk/wp-content/uploads/2021/09/The-State-of-Free-Speech-Online-1.pdf.

	harm in a way that could have a bearing on the free expression rights of survivors of self-harm and other mental illnesses. Ofcom's approach to recovery content is one which does not give adequate attention to the careful balance required to protect users' free expression rights. The proposals acknowledge that "harmful suicide and self-harm content can manifest online in various forms, ranging from recovery content that could benefit some users but be harmful to others depending on the context and individual" (Vol 3, para 7.2). However, the proposals recommend that this content can nonetheless, unwittingly, cause harm and children should be prevented from engaging with this (Vol 3, para 7.2.6). This approach will necessarily lead to the removal of the majority of recovery content, out of an abundance of caution. It is unlikely that mass, automated content moderation would be able to handle the nuanced distinction between recovery content and self-harm or suicide content. The proposals suffer from the misconception that exposure to mental illness is contagious (see Vol. 3, para 7.2.31), which Dr. Vic Baines, a former Visiting Associate at the Oxford Internet Institute disputes.[4] A shrinking private sphere may deter people from seeking social support and a safe space to freely express themselves. It is important that the internet remains a rich resource for people to openly explore mental health issues, with their rights to privately and freely access information protected. Evidence suggests that teens who suffer from mental health problems often retreat to social media, rather than the other way around.[5] The proposal acknowledges that "children with mental health conditions are significantly more likely to encounter suicide or self-harm content," which suggests that Ofcom is aware of this correlation, but seems to, perhaps wrongly, attribute it to causation (Vol 3, para 7.2.60).

4 Big Brother Watch, The State of Free Speech Online, https://bigbrotherwatch.org.uk/wp-content/uploads/2021/09/The-State-of-Free-Speech-Online-1.pdf, p95.

5 Heffer, T., Good, M., Daly, O., MacDonell, E., & Willoughby, T. (2019). The Longitudinal Association Between Social-Media Use and Depressive Symptoms Among Adolescents and Young Adults: An Empirical Reply to Twenge et al. (2018). Clinical Psychological Science, 7(3), 462-470. https://doi.org/10.1177/2167702618812727

	Question 7 The Online Safety Act does not adequately define "harm" or "harmful content" regarding either children or adults and Ofcom's ability to undertake its functions as an Online Safety regulator will suffer for this. As well as "primary priority content that is harmful to children" and "priority content that is harmful to children" the legislation and Ofcom's consultation document also identifies a category of content called "non-designated content". However, the definition attributed to this category of content is essentially meaningless, namely that undesignated content of this kind is that which presents "a material risk of significant harm to an appreciable number of UK children". Ofcom is bound by the breadth of this definition which is poorly worded, overly broad and as such presents threats to free expression. Attempts have been made to give further clarity to this definition and in the document Ofcom states that non-designated content will only be classified as such if it meets the four stage test set out at 7.9.16 of Ofcom's proposals. Despite Ofcom's attempt to provide further clarity, we are nonetheless concerned by how permissively the test has been drawn. Under strand 3 of the test, there must be some evidence to indicate a relationship between "significant harm" and a specific kind of content present on services (Vol. 3, para 7.9.25). The proposals acknowledge that it is rare to have evidence which establishes such a relationship and therefore indicate that Ofcom will rely on "research with children, parents and carers, and specialists." Given the implications of designating new categories of content harmful to children, we are not satisfied that this proposed qualitative data will be sufficient to categorise NDC. When asserting which types of content will be categorised as "non-designated content that presents a material risk of harm to children", the Act does not set a numerical threshold to determine whether an 'appreciable' number of children are facing material risk of significant harm and nor does Ofcom. The proposals suggest that Ofcom will consider the proportion of

	children encountering a given type of content once it is identified as NDC, taking into account the proportion of children with vulnerabilities who may be affected (7.9.29 and 7.9.31). This is a highly permissive test, which could result in anything that might affect a vulnerable group being designated as harmful. As the consultation notes, children with mental health conditions make up a sizeable proportion of children overall. Therefore, if Ofcom decides that a particular type of NDC may pose a threat to those children, it could be designated and censored as a result, without looking at the granularity of the issue.
Draft Guidance on Content Harmful to Children (Section 8)	
12. Do you agree with our proposed approach, including the level of specificity of examples given and the proposal to include contextual information for services to consider? 13. Do you have further evidence that can support the guidance provided on different kinds of content harmful to children? 14. For each of the harms discussed, are there additional categories of content that Ofcom a) should consider to be harmful or b) consider not to be harmful or c) where our current proposals should be reconsidered?	Confidential? – ¥ / N Question 12 Specificity and tightly-prescribed rules are always important when it comes to protecting free speech, but by virtue of the subjectivity of what constitutes harmful speech, which may be lawful expression, Ofcom may risk failing in its obligations to protect individuals Article 10 right to freedom of expression. Questions 13 and 14c Whilst it is crucial to protect people with protected characteristics from targeted hate and abuse, restrictions on broadly defined "hate speech" can impact lawful speech, if it is not properly defined. As previously mentioned, the UK has a array of criminal offences which restrict certain types of speech where it is targeted at particular groups. The categories of example abuse and hate content that is harmful to children in the proposals however, are incredibly broadly defined. They include "insulting or intimidating remarks or harmful

	stereotypes targeted towards an individual" and content "repeating harmful and discriminatory ideas about another group" (Vol. 3, Table 8.6.2). Our report, *The State of Free Speech Online*, revealed how content which engages 'harmful stereotypes' can be inappropriately censored. For example, women on Twitter have had their accounts temporarily limited, suspended or banned and tweets removed for posting that men are more likely to commit violent crime than women – despite this assertion being supported by ONS data.[6] In another example, a woman had her Twitter account suspended for 7 days after posting that the definition of rape in UK law meant that women could not perpetrate the crime.[7] Ofcom's guidance also states that 'a post that uses gendered and/or homophobic slurs to mock or degrade a person' would fall within the definition of content that is abusive and harmful to children. Our research has highlighted how words that form part of the political lexicon and are central to taxonomy of gender in some discourses on gender identities, such as "TERF" and "cis", could fall into Ofcom's category of 'gendered slurs.'[8] For example, a trans journalist's Twitter account was temporarily limited after she said "You're cis" and another user was suspended after they were reported for referring to a male user as "cis".[9] The proposals on misgendering are

6 Big Brother Watch, (n 3), pp50-51.
7 Ibid, p53.
8 Ibid, p57.
9 Ibid, pp58-59.

	likely to rely on self-reporting by users, as services have no way of knowing how individuals identify. Our evidence shows that if large social media platforms opt not to age gate their sites and as a result must moderate in a way to account for Ofcom's child safety codes, this categorisation could be exploited to censor and silence others. Further, the guidance on "insulting or intimidating remarks or harmful stereotypes" does not take into account the potentially positive effects of in-group linguistic reappropriation. This is where a group reclaims certain pejorative words or phrases that were previously directed at the group as terms of abuse. In-group linguistic reappropriation can be a source of empowerment for minority or marginalised groups, which allows for the nullification of previously harmful meanings, and is also employed by comedians from minority backgrounds to satirise otherwise offensive terms. Our research has shown that content moderation faces difficulties distinguishing between content with this level of nuance. For instance, an LGBT Instagram page was forced to reissue a post which had previously contained the word "dyke" in the caption.[10] The term, historically used as a homophobic slur to describe lesbians, has been reclaimed by the LGBT community and can be used jovially, satirically or even as a term of empowerment.

10 Ibid, p34.

	Ofcom also includes "a comment that intentionally misgenders a person with the intention to humiliate, insult, offend or 'out' someone" within the category of content that is abusive and harmful to children. Misgendering is a highly sensitive topic and can be very hurtful to trans people. However, misgendering is not a criminal offence in the UK. Indeed, compelling speech (e.g. the use of certain pronouns) could raise tensions around freedom of expression. It is possible that misgendering could be involved in crimes against transgender individuals, but a criminal offence (such as harassment or a communications offence) would have to be involved. Our investigations have revealed many examples of negative consequences associated with such a well-intended policy, including undue censorship of women and transgender people. Some unusual but very high profile cases involving gender transition demonstrate the complexities of censoring "misgendering" – for example, the survivor and estranged wife of serial rapist Isla Bryson (formerly named Adam Bryson), continues to refer to Bryson with male pronouns,[11] as did a number of mainstream media outlets. Whilst Ofcom's guidance specifies that only intentional misgendering should be classified as harmful content, this places the power primarily in the subjective interpretation of the reporter or content moderator. Granular and politicised censorship can be counter-productive and in-

11 https://www.bbc.co.uk/news/uk-scotland-64796926

	flame debates, leading to greater controversy. The guidance also identifies "content which objectifies and demeans a person on the basis of their listed characteristic" as a an example of content that is abusive and harmful to children. This includes, "a post that claims an individual is physically or mentally inferior or deficient on the basis of one or more listed characteristic(s)" and a "derogatory meme or caricature of a person, with threatening, abusive, hurtful or harmful commentary added." Our research has found examples where similarly restrictive policies resulted in over-stringent moderation. For instance, in the wake of the #MeToo movement in Autumn 2017, comic Marcia Belsky was banned from Facebook for 30 days for writing "men are scum" on a woman's post about misogyny she had experienced.[12] A tide of subsequent posts by women echoing the language were removed and enforcement action was taken against the posters. In a similar vein, a male user was banned from Facebook Messenger for seven days after joking that boys stink and a separate user breached the platform's community standards by posting, "Please let women run the world. Men are idiots."[13] Under Ofcom's guidance, it appears that similar light-hearted and lawful comments could be censored as content which is abusive and harmful to children should a regulated service be obliged to fulfil these obligations.

12 Ibid, p16.
13 Ibid, p17.

	Ofcom also includes examples of content which incites hatred, including "a comment which justifies or promotes the social exclusion of a group from society that share a listed characteristic" (Vol. 3, Table 8.6.3). Given that the listed characteristics include sex, we are concerned that this could extend to exclusionary spaces for specific protected groups (i.e., single sex spaces and services), meaning that calling for exclusion of men from an online space (e.g. for female rape survivors), and vice versa, could be categorised as inciting hatred. We welcome the inclusion of examples of content Ofcom considers not to be self-injury content at Table 8.4.3. Our investigations have revealed that many posts legitimately documenting content relating to self-harm in any way, including recovery, empowerment and healing, are removed. Furthermore, some posts that are not about self-harm at all but that contain a photo of an individual with scars were found to be obscured and marked as "sensitive."[14] The youth global network It Matters, noted that the removal of images related to self-harm by social media companies has "fuelled more stigma against mental illness" and harmed young people who reported "feelings of being 'bullied', 'isolated' and 'humiliated' by...censorship."[15] Censoring content that discusses self-harm and recovery is likely to have a chilling effect on others who have had similar experiences and are

14 Ibid, p36.

15 Ibid, p41.

	made to feel unable to share their personal photos or stories online. Far from creating an inclusive, welcoming environment for people who have experienced trauma or mental ill health, the platform actively discriminates against them. We therefore reiterate the importance of emphasising to services that not all-content relating to self-harm is harmful to children.
Volume 4: How should services assess the risk of online harms? **Governance and Accountability (Section 11)**	
15. Do you agree with the proposed governance measures to be included in the Children's Safety Codes? a) Please confirm which proposed measure your views relate to and explain your views and provide any arguments and supporting evidence. b) If you responded to our Illegal Harms Consultation and this is relevant to your response here, please signpost to the relevant parts of your prior response. 16. Do you agree with our assumption that the proposed governance measures for Children's Safety Codes could be implemented through the same process as the equivalent draft Illegal Content Codes?	Confidential? – ¥ / N It is without doubt that greater levels of accountability and transparency from online intermediaries are needed. However, we are concerned that Measure GA2, which requires service providers to name a person accountable for compliance with the children's safety duties, would guarantee widespread censorship online. Making one individual responsible for having to justify the entire services' compliance decisions, coupled with broad definitions and a low threshold of acceptable expression, will result in platforms unscrupulously removing lawful content on their sites.
Volume 5 – What should services do to mitigate the risk of online harms **Our proposals for the Children's Safety Codes (Section 13)**	

Proposed measures 22. Do you agree with our proposed package of measures for the first Children's Safety Codes? a) If not, please explain why. **Evidence gathering for future work.** 23. Do you currently employ measures or have additional evidence in the areas we have set out for future consideration? a) If so, please provide evidence of the impact, effectiveness and cost of such measures, including any results from trialling or testing of measures. 24. Are there other areas in which we should consider potential future measures for the Children's Safety Codes? a) If so, please explain why and provide supporting evidence.	Confidential? – Ɏ / N Question 22 As an organisation that seeks to uphold, promote and protect the right to freedom of expression, with a particular focus on technology, Big Brother Watch has expressed serious concerns about the Government's introduction of the Online Safety Act ('OSA') since 2019, when the Online Harms White Paper was published.[16] In the course of the OSA's passage as a Bill through Parliament we highlighted the significant implications of the proposed regulatory framework for freedom of expression and the right to privacy online and believe the Act's requirements for online platforms to surveil and restrict online speech will do significant damage to the free flow of information and ideas that the internet has facilitated. The OSA is a fundamentally flawed piece of legislation. The proposals set out in the Act, and by extension, Ofcom's Codes of Practice, will force social media companies to act as privatised speech police and will compel online intermediaries to over-remove content. The general effect of creating and enforcing codes of practice will be to fortify social media companies' terms of use, ensuring that they are upheld, and to clearly identify companies that fail to comply, who risk sanction. This new regulatory framework, which effectively amounts to overseeing private companies upholding

16 Big Brother Watch's response to the Online Harms White Paper Consultation – Big Brother Watch, July 2019: https://bigbrotherwatch.org.uk/wp-content/uploads/2020/02/Big-Brother-Watch-consultation-response-on-The-Online-Harms-White-Paper-July-2019.pdf

	those terms and conditions – sets of rules that are not neutral and which have complex human rights and data protection implications - will pose threats to free expression and privacy in the UK. The UK already has expansive laws governing speech-related offences that can be used to prosecute violent, hateful and harmful forms of speech and behaviour online. This includes laws prohibiting speech that causes harassment, alarm, distress, or fear (Protection from Harassment Act 1997; Public Order Act 1986); speech that is deemed grossly offensive and purposefully annoying or distressing (Malicious Communications Act 1988; Communications Act 2003); and speech that incites hatred on the basis of race, religion or sexual orientation (Crime and Disorder Act 1998; Race and Religious Hatred Act 2006). It remains our view that law enforcement agencies could better use these laws to deal with many of the harms children might experience online in collaboration with the largest social media companies. Instead, the OSA and Ofcom's proposals will see these private companies deputised by the state to act as private online law enforcement bodies, tasked with restricting children's rights to freely impart and receive information, far beyond pre-existing legal boundaries, which in our view will lead to a new wave of privatised monitoring and censorship. In Vol. 5, section 13, Ofcom sets out five broad areas within which its proposals fall, including robust age checks, safer algorithms, effective

	moderation, strong governance and accountability and more choice and support for children. Measures RS1 and RS2 propose that services use recommender systems to exclude content likely to be Primary Priority Content ("PPC") and limit the prominence of content likely to be Priority Content ("PC") for children (Vol. 5, para 13.29). Whilst we agree that recommender systems can place children in harm's way online, we believe this is an argument for the dismantling of the platform's business models and the mass data collection practices which inform these algorithmic feedback loops, rather than trying to use tools to prevent certain types of expression being recommended. Further, we disagree that blanket-filtering out all content that might be harmful to children is the best approach to keeping children safe online (Vol. 5, para 13.32). Whilst some content that is harmful to children will always be clear and obvious to content moderators, it is difficult to see how they should be able to make determinations on nuanced content, including recovery content, and what constitutes "insulting or intimidating remarks" or "repeating harmful and discriminatory ideas about another group," which is often highly contextual. Expecting Silicon Valley's content moderators to undertake these complex decisions at speed, without accountability and under the threat of penalities, will almost certainly lead to the over-censorship of speech from those platforms who will have to observe these guidelines out of an abundance of caution.

	We welcome Ofcom's decision to not mandate the use of automated tools for general content moderation (Vol. 5, para 13.34), although we remain concerned that given the legal burden placed on services to moderate content, many will inevitably have no choice but to use automated tools to fulfil their obligations. The use of automated tools for content moderation necessitates the mass scanning and automated analysis of all online content, which often results in the surveillance and over-removal of online expression given the limitations of the technology to detect nuance (which Ofcom acknowledges at Vol. 5, para 13.64) as well as a wider chilling effect on user's speech (see our response to Question 36 for further detail). Section 12(3)(a) of the Act sets out platforms' need to "prevent" children from encountering PPC that is harmful, however this approach to content moderation is one which poses serious threats to freedom of expression. In order to truly "prevent" illegal or "harmful" content, platforms would have had to pre-screen content through upload filters. This was described by internet lawyer, Graham Smith, as having a "predictive policing element"[17] and as Dan Squires KC and Emma Foubister have argued in a legal opinion commissioned by Open Rights Group, this would be a form of prior restraint which is a serious violation of the right to freedom of speech.[18] Ofcom's approach, which recommends removing lawful online

17 Smith, G. Mapping the Online Safety Bill, Cyberleagle blog, 27 March, 2022 https://www.cyberleagle.com/

18 Dan Squires KC and Emma Foubister, In the Matter of: The Prior Restraint Provisions in the Online Safety Bill, Matrix Chambers, Commissioned by Open Rights Group, https://www.openrightsgroup.org/publications/legal-advice-on-prior-restraint-provisions-in-the-online-safety-bill/

	content, is an approach which does not align with the legal standards previously held by liberal democracies when it comes to the regulation of online expression and, in this regard, is a dangerous approach that risks failing Ofcom's obligations under the HRA. Ofcom adopts a contradictory approach in relation to the treatment of children of different ages, on the one hand "encourag[ing] services to tailor their experiences to children in different age groups," whilst acknowledging that there are "limited existing technologies that can reliably identify children of different ages" (Vol. 5, para 13.75). Ofcom's guidance suggests that services will be able to do this "based on their understanding of their user base," however it is unlikely that providers, in particular smaller services, will have access to this kind of granular information. Given that services will not know how old individual users are, tailoring children's online experience for their age group is an impossibility. It is important to note that children have free expression rights under the EU Charter and Convention on the Rights of the Child, which entitle them to "seek, receive and impart information and ideas of all kinds."[19] Infringing these rights to protect younger children from harm, simply because of technological limitations, cannot be proportionate. We welcome the recommendation that the HEAA obligations will no longer fall on search

19 Article 13, Convention on the Rights of the Child 1989

	services (Vol. 5, para 15.8). Given the educational benefits of Google and its wide use across all sections of society, including children, it is important that the search giant need not adopt invasive identification policies or else sanitise the information readily available to the general public.

Developing the Children's Safety Codes: Our framework (Section 14)	
25. Do you agree with our approach to developing the proposed measures for the Children's Safety Codes? a) If not, please explain why. 26. Do you agree with our approach and proposed changes to the draft Illegal Content Codes to further protect children and accommodate for potential synergies in how systems and processes manage both content harmful to children and illegal content? a) Please explain your views. 27. Do you agree that most measures should apply to services that are either large services or smaller services that present a medium or high level of risk to children? 28. Do you agree with our definition of 'large' and with how we apply this in our recommendations? 29. Do you agree with our definition of 'multi-risk' and with how we apply this in our recommendations? 30. Do you agree with the proposed measures that we recommend for all services, even those that are small and low-risk?	Confidential? – Ɏ / N Question 25 The Act compels platforms to "prevent" children from encountering content of this kind which will lead to greater levels of monitoring and surveilling users' activity in order to fulfil these preventative policing-style obligations. The result of the proposals would be that unless a platform undertakes invasive age verification checks and then age-gates user-generated content at a granular level, content moderation on the site in question must be tailored for children. Requiring all adults to verify they are over 18 in order to access everyday online services is a disproportionate response to the aim of protecting children online and violates fundamental rights. It carries significant risks of tracking, data breaches and fraud by mandating users to volunteer even more personal information to private platforms, which could be stored in large centralised databases. It undermines anonymity, which is widely recognised as a vital option by human rights experts for the full enjoyment of rights and equality online, and it

	creates digital exclusion for individuals unable or unwilling to meet requirements to show formal identification documents. Where age-gating also means the collection of children's data en-masse, the privacy risks are magnified. These obligations will also put an onerous burden on small-to-medium enterprises, and as such will ultimately entrench the market dominance of large tech companies and lessen choice and agency for both children and adults. We remain concerned by the impact that proposals to moderate search engine content will have on freedom of expression and access to information online. The right to freedom of expression in an online setting not only concerns the ability of individuals to impart information but also to receive it. In this regard, a free flow of information and the right to freedom of expression go hand in hand. Question 26 We welcome the proposed additional measures for the Illegal Content Codes where they are designed to improve transparency, accessibility and user choice and control. Question 30 We do not agree that small and low-risk services should employ the terms of service and content moderation processes requirements (see our responses to Questions 36, 38 and 46 for further detail). The proposals acknowledge that the cost implications could cause smaller services to stop serving adult users in the UK (Vol. 5, para 15.119). This would be an unac-

	ceptable encroachment on the free expression rights of these services' users, who should be able to access information from a plurality of sources.
Age assurance measures (Section 15)	
31. Do you agree with our proposal to recommend the use of highly effective age assurance to support Measures AA1-6? Please provide any information or evidence to support your views. a) Are there any cases in which HEAA may not be appropriate and proportionate? b) In this case, are there alternative approaches to age assurance which would be better suited? 32. Do you agree with the scope of the services captured by AA1-6? 33. Do you have any information or evidence on different ways that services could use highly effective age assurance to meet the outcome that children are prevented from encountering identified PPC, or protected from encountering identified PC under Measures AA3 and AA4, respectively? 34. Do you have any comments on our assessment of the implications of the proposed Measures AA1-6 on children, adults or services? a) Please provide any supporting information or evidence in support of	Confidential? – Y / N Question 31 We have serious concerns about the proposals requiring widespread age verification across websites, apps and other online services, particularly at the point of entry, which will lead to increased data profiling of both children and adults, and restrictions on their freedom of expression and access to information. Firstly, the efficacy of these systems has been disputed. Ofcom Chief Executive Melanie Dawes said "age assurance technologies which scan your face and estimate your age don't work very well on children because children can look so different at different ages."[20] A position paper published by 20 European civil society organisations last year, reiterated the absence of evidence that measures like document-based age verification and age estimation are effective at sheltering children from online harm.[21] The same paper also highlights the security risks involved in the mass collection and processing of users' personal data, which could

20 https://www.biometricupdate.com/202406/yoti-responds-to-ofcoms-counterfactual-statements-on-age-assurance-tech

21 EDRi, Position Paper: Online age verification and children's rights, 4 October 2023, https://edri.org/wp-content/uploads/2023/10/Online-age-verification-and-childrens-rights-EDRi-position-paper.pdf

your views. 35. Do you have any information or evidence on other ways that services could consider different age groups when using age assurance to protect children in age groups judged to be at risk of harm from encountering PC?	put children at risk of harm in the event of a security breach. The Electronic Frontier Foundation emphasised that "once information is shared to verify age, there's no way for a website visitor to be certain that the data they're handing over is not going to be retained and used by the website, or further shared or even sold."[22] Data protection legislation will not necessarily afford individuals sufficient protections to guard against this threat, especially given the mandatory nature of the processing in order to access services. Finally, the implementation of HEAA requirements will create digital exclusion for individuals unable to meet requirements to show formal identification documents. AA2 The OSA only requires services to *protect* children in age groups judged to be at risk of harm from PC content. Ofcom is therefore "exercising a degree of discretion by recommending the use of highly effective age assurance in relation to identified PC" at the point of entry under measure AA2 (Vol. 5, para 15.105). Such a measure would *prevent* children from encountering PC content and is an unnecessary and disproportionate measure. Although the Act only requires services to protect children of an age that is judged to be at risk of harm, Ofcom does not recommend that the measure be tailored to particular age groups, due to limited

22 Electronic Frontier Foundation, Age Verification Mandates Would Undermine Anonymity Online, 10 March 2023, https://www.eff.org/deeplinks/2023/03/age-verification-mandates-would-undermine-anonymity-online

	evidence on the technical capability of distinguishing between age groups. This indicates that HEAA technology is not appropriate to deal with the specific harm set out in the Act and its wholesale application of content filtering obligations to services under measure AA2 as an access control would be disproportionate and have a major impact on users' rights. <u>AA3</u> Ofcom recommends applying HEAA to prevent children's access to PPC on services whose principal purpose is not the hosting or dissemination of PPC, but which do not otherwise prohibit PPC content (Measure AA3). The Regulator acknowledges both that such services will "host a significant amount of non-harmful content" and that age assurance processes are just "one way" of preventing children from accessing PPC content on these services (Vol. 5, para 15.133 and 15.134). Providers of such services are permitted to exercise discretion over where to position the age assurance process on their service to prevent children from encountering PPC (Vol. 5, para 15.134). Given that the amount of PPC content on these services is unlikely to be high, as it is not their principal purpose, preventing children from accessing the entire service would be disproportionate. Any use of any age assurance technology should be implemented at the least restrictive point of access to the service. In other words, users should be able to engage with the service as much as possible and only asked to verify their age when engaging with the content in question, e.g. pornography. Further, this can only be implemented with

	clear data protection safeguards in place which surpass the basic data protection principles of GDPR. There is also a real risk that despite providers having discretion over what stage HEAA can be implemented, some services might choose to restrict access to all children for cost or complexity reasons (Vol. 5, paras 15.148 and 15.151). This has worrying implications for the free expression rights of children who will be prevented from seeing vast amounts of non-harmful content and for adults who will be forced to engage in highly intrusive data HEAA data processing to access content which, in any event, is largely unharmful to children. The measure could alternatively lead these types of services to change their terms of service to prohibit all forms of PPC in order to avoid the costs associated with compliance (Vol. 5, para 15.153). This would have similarly harmful consequences for adult users, as they would no longer be able to access content which is legal, but falls under the PC category of content that is harmful to children. We recognise the need to regulate pornographic content and to do so in a way which prevents children from accessing material of this kind – as attempted under measure AA1. This should be focused on those services whose primary purpose is the creation or hosting of such content. However, as per the highlighted risks set out above, such a scheme cannot proceed without embedding serious privacy safeguards in its application. The collection of identity documents or biometric data

	for access to adult-content websites is a recipe for disaster, matching personal identifiers with adults' sensitive viewing habits. Not only does this risk compromising intimate elements of individuals' private lives but it poses a threat to members of the LGBT community who may not be "out" and openly willing to reveal their sexual preferences. Ofcom points to data protection legislation as a safeguard for users' privacy rights (Vol. 5, para 15.77), however, as Open Rights Group have observed, GDPR has provided a number of safeguards when it comes to data protection, but it does not, on its own, protect information that is as potentially revealing as a person's pornographic viewing history.[23] The organisation has set out other minimum standards for achieving a system which is safe and secure and argues that in order to safeguard individuals' privacy age verification systems should: "process the minimum personal data necessary to verify your age; additional personal data should not be collected, irrespective of whether it is subsequently securely deleted. Personal data must not be kept for longer than is necessary to achieve the purpose of age verification, and must not be used for other purposes, such as advertising." Similar provisions set out in the Digital Economy Act, which related to age-verification checks for websites which hosted pornographic content, delegated responsibility for this area to the British Board of Film Classifica-

23 Open Rights Group, Age Verficiation Facts, https://www.ageverificationfacts.org.uk/over-18s/

	tion (BBFC) in line with their other responsibilities to regulate in this area. However, the BBFC's certification scheme for providers of age-verification technologies was voluntary, which would have resulted in non-secure providers using this new compelled system to harvest individuals' most sensitive personal data. Unless these problems are addressed, the systems Ofcom introduce for age-assurance relating to services whose primary purpose is the creation or hosting of pornographic content will suffer from the same flaws and will create inherent privacy risks for adults online. The same principles apply to measure AA4, though this proposal is even more disproportionate given the narrower requirement under the Act to *protect* children in age groups judged to be at risk of harm from PC content. Question 34 Ofcom acknowledges that its measures could result in potentially significant impacts on adult users' ability to access the service," including services who cannot afford the cost of implementing age assurance exiting the UK (15.66-15.67). However, it suggests that since the bigger services would be able to absorb the cost, users could simply switch to these available services. This downplays the extent to which this would harm the rights of adult users to access information. Market concentration would result in less choice for users over where to receive and impart information in line with their free expression rights. The proposals also acknowledge that the measures will make it more cumbersome for

	adults to access these services, and the way services implement age assurance could in some cases dissuade adult users from using the service altogether (Vol. 5, para 15.68). For instance, users might be dissuaded by a one-off age assurance check or having to complete age assurance checks each time they use the site, due to the onerousness of the measures or privacy concerns. Ofcom concludes that the impact on users' freedom of expression and association would be "relatively limited" in this scenario as they are choosing not to use the age assurance mechanism. However, given it is the only way they can gain access to the service, any consent cannot properly be understood to be freely given. Although "providers have incentives to make their age assurance process as user-friendly as possible and limit friction to adult users" this does not address the privacy concerns that some users will still have. Ofcom notes that "it is possible to assure a user's age without retaining data other than as needed for the purposes of the age check" (Vol. 5, para 15.76), however there are no safeguards proposed to ensure this is the case. Although Ofcom does not require providers to rely on ID, the other proposed forms of HEAA also necessitate the processing of highly sensitive biometric data, with questions over its accuracy and data retention by third parties. Even whilst adhering to the principle of data minimisation (Vol. 5, para 15.77), this would give private companies, many of which have a concerning track record of data exploitation, an extraordinary amount of personal information about their users, linking a user's

	online activity to their offline identity. Ofcom also noted that some adult users may be mistakenly identified as children and restricted from using a service. Although the proposals advocate giving the user a mechanism for redress, this will only be useful retrospectively and in the meantime, will pose an impossible restriction to those who have been misidentified whilst trying to exercise their free expression rights.
Content moderation U2U (Section 16)	
36. Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views. 37. Do you agree with the proposed addition of Measure 4G to the Illegal Content Codes? a) Please provide any arguments and supporting evidence.	Confidential? – ¥ / N Automation is a blunt tool for content moderation, which deals with nuanced areas of speech, law and the adjudication of individuals' rights. Whilst automation can play a role in detecting the most serious illegal material, the use of such tools should be strictly limited, and certainly should not extend to posts which could be classified as helpful 'recovery' content or the 'incitement of hatred'. AI-powered content moderation systems often censor speech that is both awful and compliant with platforms' terms of use.[24] Big Brother Watch has extensively documented examples of major platforms removing lawful speech which has been wrongly flagged as 'hate'. Topics as varied as gender identity, criticism of police racism, jokes about gender stereotypes, sexuality, and statistics about crime have all been flagged by

24 Big Brother Watch, State of Free Speech Online, September 2021, https://bigbrotherwatch.org.uk/wp-content/uploads/2021/09/The-State-of-Free-Speech-Online-1.pdf.

	platforms as inciting hatred and removed (see our response to Q13 for further detail). Whilst a cautious censorial approach may be seen as a worthwhile trade-off for a commercial platform seeking to sanitise content for advertisers and maximise profits rather than users' rights, Ofcom is a public body subject to obligations under the HRA and must consider the risks to free expression of censorial measures. Under rigorous obligations to protect children from harmful content on their sites, online intermediaries will either lock users out of their sites altogether with age-gating or over-remove content on their platforms under the threat of penalties. The consequential impact on free speech will be profound. We welcome Ofcom's decision to not mandate the use of automated tools for general content moderation (Vol. 5, para 13.34), although we remain concerned that given the legal burden placed on services to moderate content, many will inevitably have no choice but to use automated tools to fulfil their obligations. The use of automated tools for content moderation necessitates the mass scanning and automated analysis of all online content, which often results in over-removal of online expression given the limitations of the technology to detect nuance (which Ofcom acknowledges at Vol. 5, para 13.64) as well as a wider chilling effect on users' speech and significant privacy intrusion for all users (see our response to Question 36 for further detail). Ofcom's contradictory recommendation at paragraph 16.16 that services "that are not

	currently deploying automated technologies for content detection [should] invest in systems that will help detect this content in their services at scale," is, for this reason, troubling. We have doubts about the premise that simply more content moderation from social media sites will keep children safe online. Ofcom notes that content moderation systems and processes are already employed by a number of services that are likely to be accessed by children (Vol. 5, para 16.41) but these systems have limitations. The Register of Risk acknowledges the difficulty with identifying content where it is deliberately obscured to prevent detection by content moderation systems. It is not clear that requiring all U2U services to fortify these systems would prevent users from circumventing keyword detection, as has historically been the case. Under the obligations set out by the OSA and Ofcom, platforms will naturally automate more of their content moderation systems. However, this is not appropriate for many of the specific types of content the Act is designed to target. Our research shows that content moderation is unable to distinguish between self-harm content and recovery content.[25] Similarly, examples given for bullying content are content that "persistently or repetitively targets individuals or groups with offensive or otherwise harmful content," (Vol. 5, para 7.5.5). It is difficult to see how automated or semi-automated moderation

25 Ibid, pp37-39.

	would be able to identify such information, unless it tracked and logged users' behaviour over time. Further, having social media companies simply take-down content which is deemed to be "harmful" does nothing to tackle the root source of the problem. Even the provisions enabling services to suspend functionality or ban and suspend users only stifle speech but do not address the behaviour of individuals who could simply open new accounts and repeat the same behaviour. Where content or behaviour is manifestly illegal, it should be dealt with by legal enforcement, not by social media companies who are merely able to obscure content. We welcome Ofcom's decision under Measure 4A not to require pre-publication moderation of content, instead requiring services to "swiftly action" content. In order to truly "prevent" children from encountering PPC content, platforms would have had to pre-screen content through upload filters. This was described by internet lawyer, Graham Smith, as having a "predictive policing element"[26] and as Dan Squires KC and Emma Foubister have argued in a legal opinion commissioned by Open Rights Group, this would be a form of prior restraint which is a serious violation of the right to freedom of speech.[27] We are concerned by the contradictory proposal at Vol.

26 Smith, G. Mapping the Online Safety Bill, Cyberleagle blog, 27 March, 2022, https://www.cyberleagle.com/

27 Dan Squires KC and Emma Foubister, In the matter of: the prior restraint provisions in the Online Safety Bill, Matrix Chambers, Commissioned by Open Rights Group, https://www.openrightsgroup.org/publications/legal-advice-on-prior-restraint-provisions-in-the-online-safety-bill/

	5, para 16.34, in relation to services whose primary purpose is not the hosting or dissemination of PPC but that do not prohibit all kinds of PPC in their terms of service. Ofcom recommends that for such services, providers should "take appropriate action such as using filtering – so that each piece of content identified as PPC is only visible to users confirmed to be adults using highly effective age assurance – or ensuring that all identified PPC is present only on parts of the service where access is restricted to users confirmed to be adults using highly effective age assurance." It is unclear whether this encourages the prescreening of content, which for the reasons aforementioned, we would strongly oppose. We would urge Ofcom to provide clarity on this point. Ofcom indicates that its proposed measure under 4A is not limited only to content or communications that are communicated publicly, and may lead to the review of content or communications in relation to which individuals might expect a reasonable degree of privacy (Vol. 5, para 16.57). This could involve surveillance of private messages and even the use of a technique used to circumvent end-to-end encrypted messaging services at scale, known as client-side scanning ('CSS'), which would create vulnerabilities within messaging services for criminals to exploit or could open the door to a greater level of mass surveillance through use of this technology.[28]

28 Fact Sheet: Client-Side Scanning, The Internet Society, March 2021, https://www.internetsociety.org/resources/doc/2020/fact-sheet-client-side-scanning/

	Tackling content that is harmful for children does not require entire encrypted channels to be compromised, sacrificing the security, safety and privacy of billions of people. Wherever surveillance is carried out, it should be targeted and based on suspicion in line with UK law. In a legal opinion commissioned by the free expression organisation, Index on Censorship, Matthew Ryder KC and Aidan Wills of Matrix Chambers found that mandating the general screening of users' private communications through technology such as CSS would be a disproportionate interference with the rights to privacy and freedom of expression unless the state is "confronted with a serious threat to national security which is shown to be genuine and present or foreseeable" (and other criteria are satisfied) (La Quadrature; Ekimdzhiev v Bulgaria (2022) 75 EHRR 8, [138] – [139], [168]).[29] The surveillance of millions of lawful users of private messaging apps has been found to require an extremely high threshold of legal justification, which generalised content moderation purposes would be highly unlikely to meet. Currently, this level of mass scale, state mandated surveillance would only be possible under the Investigatory Powers Act if there is a credible threat to national security. Ofcom should not mandate the use of CSS for any purposes under the Online Safety Act. Measure 4B proposes that services should set clear internal content moderation policies,

29 Surveilled and Exposed: How the Online Safety Bill Creates Insecurity – Index on Censorship, November 2022: https://indexoncensorship.org/wp-content/uploads/2022/11/Surveilled-Exposed-Index-on-Censorship-report-Nov-2022.pdf

	because publishing information about policies publicly can allow users to circumvent the content moderation systems and processes (Vol. 5, para 16.85). Given the extensive power these private companies have to moderate speech online, this would establish a deeply concerning precedent, whereby users are kept in the dark about how and why their speech is assessed and removed. No such approach would ever be taken with regards to censorship under British law, as it contradicts the principles of accessibility and foreseeability under the rule of law. It is vital that platforms operate in the spirit of transparency and create acessible rules on their sites which respect human rights standards. We appreciate Ofcom's recommendation that "when setting targets, services should balance the desirability of swiftly actioning content against the desirability of making accurate moderation decisions" under CM3 (Vol. 5, para 16.114). Nevertheless, we remain alarmed by Ofcom's requirement that companies set targets for content moderation. While the requirement that companies assess the accuracy of their content moderation is welcome, we are concerned that setting targets for the time taken to remove content will pressure companies, and individual workers dealing with challenging information, to remove content at pace. Any pressure on moderators to meet certain time goals will inevitably lead to rushed decisions. Ofcom acknowledges these risks but suggests that the requirement that "that services will need to balance the speed of decisions made with the

	degree of accuracy...will mitigate the risk of unjustifiable interference with users' rights" (16.122). However, we believe that the fines and other penalties associated with non-compliance for social media companies mean that, in practice, the emphasis on removal trumps any emphasis on accuracy. Such a chilling effect has already been seen in Germany, since the Network Enforcement Act 2017 ('NetzDG') was passed. The Act threatens fines of up to €50 million for social media companies that fail to remove illegal content within 24 hours. This time frame for removal incentivises social media companies to err on the side of caution and over-censor content. Human Rights Watch has called on German lawmakers to "promptly reverse" NetzDG and explained that it is "vague, over-broad, and turns private companies into over-zealous censors to avoid steep fines, leaving users with no judicial oversight or right to appeal."[30] Similarly, Article 19 warned that "the Act will severely undermine freedom of expression in Germany, and is already setting a dangerous example to other countries that more vigorously apply criminal provisions to quash dissent and criticism, including against journalists and human rights defenders."[31] The former UN Special Rapporteur on Freedom of Expression, David Kaye, warned that NetzDG "raises serious concerns about freedom of expression and the right to privacy

30 Germany: Flawed Social Media Law – Human Rights Watch, 14 Feb 2018: https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law

31 Germany: Act to Improve Enforcement of the Law on Social Networks undermines free expression – Article 19, 1 Sept 2017, https://www.article19.org/resources/germany-act-to-improve-enforcement-of-the-law-on-social-networks- undermines-free-expression

	online", and argued that "censorship measures should not be delegated to private entities."[32] The law has also been criticised by the German broadcast media for turning controversial and censored voices into "opinion martyrs."[33] We have concerns about the factors that Ofcom lists as relevant for decisions on prioritisation of content for review under Measure 4D. For instance, the virality of a piece of content should not automatically mean that it should be subject to a higher level of scrutiny. Numerous viral challenges, such as the ALS Ice Bucket challenge, or viral videos on pressing social matters, such as the murder of George Floyd, were important for communicating a message precisely because they were widely shared. In short, content being popular should not automatically result in higher levels of surveillance. Though we welcome Ofcom's decision not to currently recommend the use of DRCs and Trusted Flaggers in this iteration of the Code, we are troubled by its recommendation under Measure 4D that services currently employing them should give priority for review to content flagged via these channels (Vol. 5, para 16.155). It is not necessarily the case that these flags are more accurate, and in some cases could lead to state authorities leaning on services to remove content they otherwise

32 Mandate of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, 1 June 2017: https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL-DEU-1-2017.pd

33 Tough new German law puts tech firms and free speech in spotlight - Philip Oltermann, The Guardian, 5 January 2018: https://www.theguardian.com/world/2018/jan/05/tough-new-german-law-puts-tech-firms-and-free-speech-in-spotlight

	would not. 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 pressured to remove content that was both lawful and not against companies' terms and conditions, raising wider concerns about the extent to which these relationships between state bodies and social media platforms are both transparent and rights-respecting.[34] When a piece of content is flagged by the state to a social media company, it places additional pressure on the company to censor the material in question. Giving state officials an unaccountable shortcut to flagging speech for removal from the digital public square poses serious threats to free speech. Not only can the government exercise its own discretion in identifying the content it thinks is objectionable and may breach terms of services, undermining the universal application of the right to freedom of speech, but this special relationship could put political content in a 'VIP' deletion lane and hasten censorship as a result. Even where flaggers are designated as bodies whose primary purpose is to ensure the safety of children online, these relationships must still be scrutinised closely to ensure human rights and civil liberties are protected. In-

34 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

	deed, it is unclear on what basis trusted flaggers would "demonstrate accuracy and reliability in flagging content" (Vol. 5, para 16.155). There is no reason why these bodies could not adopt a politicised approach to flagging content, and, given the rights implications, they too should be subject to transparency. A November 2022 review conducted by the Oversight Board, the quasi-independent "supreme court" that examines some content moderation decisions made by Meta, revealed the additional weight given by Meta to reports made by governments and law enforcement. The Oversight Board found that Meta had wrongly applied rules over "veiled threats" when it removed a drill music video by a London-based rapper, following a request from the Metropolitan Police.[35] In a lengthy ruling the Board outlined how flags from the state are handled –stating that as well as the publicly available reporting processes, requests for review from police and other arms of government are handed "at escalation" meaning they are sent to specialist internal teams at Meta, not general content moderators. In the ruling, the Board was critical of the lack of transparency and appeal rights when content moderation decisions are made "at escalation", highlighting that Meta teams often relied on evidence to justify bans from the same third parties that reported the content in the first place, including government agencies,

35 Oversight Board Overturns Meta's Decision In "UK Drill Music" Case, Oversight Board Press Release, November 2022, https://www.oversightboard.com/news/413988857616451-oversight-board-overturns-meta-s-decision-in-uk-drill-music-case/

	undermining moderators' ability to make independent judgements. The requirement to prioritise "trusted flaggers" by Ofcom gives credence and favour to a system which creates threats to human rights and at its worst enables extra-legal executive censorship.
Search moderation (Section 17)	
38. Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views. 39. Are there additional steps that services take to protect children from the harms set out in the Act? a) If so, how effective are they? 40. Regarding Measure SM2, do you agree that it is proportionate to preclude users believed to be a child from turning the safe search settings off? The use of Generative AI (GenAI), see Introduction to Volume 5, to facilitate search is an emerging development, which may include where search services have integrated GenAI into their functionalities, as well as where standalone GenAI services perform search functions. There is currently limited evidence on how the use of GenAI in search services may affect the implementation of the safety measures as set out in this code. We welcome further evidence from stakeholders on the following questions and please provider arguments and evidence to support your views: 41. Do you consider that it is technically feasible to apply the proposed code measures in respect of GenAI	Confidential? – Y / N Question 38 We remain concerned by the impact that proposals to moderate search engine content will have on freedom of expression and access to information online. The right to freedom of expression in an online setting not only concerns the ability of individuals to impart information but also to receive it. In this regard, a free flow of information and the right to freedom of expression go hand in hand. The requirement to undertake this moderation at scale will likely to lead to swathes of lawful content being erroneously downranked by search engines. It is unclear why the wording of proposed Measure SM1 that "all search services should have moderation systems and processes in place to take appropriate action on content that is harmful to children, which includes PPC, PC and NDC" should differ from Measure 1 under the draft Illegal Content Codes, which only requires services to "deindex or downrank illegal content of which it is aware, that may appear in search results." It is unclear why there is a discrepancy and Ofcom should explain this. This approach to PPC, PC

functionalities which are likely to perform or be integrated into search functions? 42. What additional search moderation measures might be applicable where GenAI performs or is integrated into search functions?	and NDC content will have significant impact on access to information, an important part of the public's right to freedom of expression and information. This is particularly disproportionate given the Act only requires services to take proportionate steps to *minimise* the risk of children encountering PPC, PC and NDC.[36] This is a far more intrusive measure to prevent children from encountering harmful content, and would result in large amounts of lawful expression being removed, downranked and de-indexed for all users, including adults. Whilst we welcome Ofcom's current assessment that requiring services to employ HEAA is not proportionate "at this stage," we are troubled by the suggestion that it is "something we may consider in the future" (Vol. 5, para 17.7). We are firmly opposed to the requirement to use HEAA for the reasons aforementioned in response to questions 31 and 34, however our opposition is particular acute in relation to search services, given their importance for accessing information. For the reasons set out in our response to question 36, we believe Measures SM3-SM6 are unnecessary and have worrying implications for users' rights. Question 40 Measure SM2 provides that "large general search services should apply a safe search setting for all users believed to be a child which filters out identified PPC from search results. Users believed to be a child should not

36 Section 29(3) of the Act.

	be able to switch this setting off" (Vol. 5, para 17.22). This proposal sets an extremely concerning test, which means that users will have to positively be proved to be adults in order to use search services (Vol. 5, para 17.38). As a result, all adult users will only be able to browse a sanitised version of the internet unless search services have reasonable grounds to believe they are an adult. Ofcom also suggests that self-declaration on sign-up, user profiling technologies, or other tools that do not amount to highly effective age assurance may be appropriate (Vol. 5, para 17.152). We consider that it is an unnecessary and disproportionate that search services should be able to censor what a user can see by guessing their age through tracking, and that it would be particularly dangerous for Ofcom to encourage companies that already trade in controversial data practices, such as Google, to undertake such profiling Ofcom suggests that unless an adult is "incorrectly determined to be a child...we do not envisage them being impacted" (Vol. 5, para 17.139). Without evidence of how effectively search services can currently identify the age of users, this is a unproven assumption. As we have previously stated in this consultation response, while it may be appropriate to determine the age of individuals where they try to access content related to a regulated industry with stringent safeguards, e.g. pornography, to apply age-gating to general search services in the way described in the consultation document is entirely disproportionate and will have a bearing on the free flow of information.

User reporting and complaints (Section 18)	
43. Do you agree with the proposed user reporting measures to be included in the draft Children's Safety Codes? a) Please confirm which proposed measure your views relate to and explain your views and provide any arguments and supporting evidence. b) If you responded to our Illegal Harms Consultation and this is relevant to your response here, please signpost to the relevant parts of your prior response. 44. Do you agree with our proposals to apply each of Measures UR2 (e) and UR3 (b) to all services likely to be accessed by children for all types of complaints? a) Please confirm which proposed measure your views relate to and explain your views and provide any arguments and supporting evidence. b) If you responded to our Illegal Harms Consultation and this is relevant to your response here, please signpost to the relevant parts of your prior response. 45. Do you agree with the inclusion of the proposed changes to Measures UR2 and UR3 in the Illegal Content Codes (Measures 5B and 5C)? a) Please provide any arguments and supporting evidence.	Confidential? – ¥ / N We would reiterate our response to question 28 of the Illegal Harms Consultation in relation to the user reporting and complaints proposals.[37]

37 Big Brother Watch Response to Online Harms Consultation, https://bigbrotherwatch.org.uk/wp-content/uploads/2024/02/Ofcom-consultation-on-illegal-harms-response-Final.pdf, p10-11.

Terms of service and publicly available statements (Section 19)	
46. Do you agree with the proposed Terms of Service / Publicly Available Statements measures to be included in the Children's Safety Codes? a) Please confirm which proposed measures your views relate to and provide any arguments and supporting evidence. b) If you responded to our illegal harms consultation and this is relevant to your response here, please signpost to the relevant parts of your prior response. 47. Can you identify any further characteristics that may improve the clarity and accessibility of terms and statements for children? 48. Do you agree with the proposed addition of Measure 6AA to the Illegal Content Codes? a) Please provide any arguments and supporting evidence.	Confidential? – Y / N We would reiterate our response to question 29 of the Illegal Harms Consultation in relation to the terms of service and publicly available statements proposals.[38]

38 Big Brother Watch Response to Online Harms Consultation, https://bigbrotherwatch.org.uk/wp-content/uploads/2024/02/Ofcom-consultation-on-illegal-harms-response-Final.pdf, p11-12.

Combined Impact Assessment (Section 23)	
58. Do you agree that our package of proposed measures is proportionate, taking into account the impact on children's safety online as well as the implications on different kinds of services?	Confidential? – Y / N As we have set out, the practical impact of Ofcom's guidance as it stands is that either all users will have to access online services via a 'sorting' age-gate or adult users will have to access the lowest common denominator version of services with an option to 'age-gate up'. This creates a de facto compulsory requirement for age-verification, which in turn puts in place a de facto restriction for both children and adults on access to online content. Requiring all adults to verify they are over 18 in order to access everyday online services is a disproportionate response to the aim of protecting children online and violates fundamental rights. It carries significant risks of tracking, data breaches and fraud. It creates digital exclusion for individuals unable to meet requirements to show formal identification documents. Where age-gating also applies to under-18s, this violation and exclusion is magnified. It will put an onerous burden on small-to-medium enterprises, which will ultimately entrench the market dominance of large tech companies and lessen choice and agency for both children and adults.
Annexes **Impact Assessments (Annex A14)**	

60. In relation to our equality impact assessment, do you agree that some of our proposals would have a positive impact on certain groups? 61. In relation to our Welsh language assessment, do you agree that our proposals are likely to have positive, or more positive impacts on opportunities to use Welsh and treating Welsh no less favourably than English? a) If you disagree, please explain why, including how you consider these proposals could be revised to have positive effects or more positive effects, or no adverse effects or fewer adverse effects on opportunities to use Welsh and treating Welsh no less favourably than English.	Confidential? – Y / N We would emphasise the impact the proposals are likely to have on marginalised groups who do not have access to identity documents. EDRi notes the "disproportionate impacts on children, people in situations of homelessness, undocumented people and other people facing social exclusion," that are associated with making identification mandatory.[39] Facial recognition age-estimation also has inherent biases which means that women, young people, and people of colour are more likely to have their rights infringed by inaccurate age-estimation. It is also important to highlight the way in which certain groups are particularly at risk to censorship arising from content moderation. For example, the removal of self-harm images on social media has been reported to fuel more stigma against mental illness and leave children feeling 'isolated' and 'humiliated' by...censorship."[40] Censoring content that discusses self-harm and recovery is likely to have a chilling effect on others who have had similar experiences and are made to feel unable to share their personal photos or stories online. Far from creating an inclusive, welcoming environment for people who have experienced trauma or mental ill health, the platform actively discriminates against them.

39 EDRi, Position Paper: Online age verification and children's rights, 4 October 2023, https://edri.org/wp-content/uploads/2023/10/Online-age-verification-and-childrens-rights-EDRi-position-paper.pdf, p5.

40 Big Brother Watch, The State of Free Speech Online, September 2021, https://bigbrother-watch.org.uk/wp-content/uploads/2021/09/The-State-of-Free-Speech-Online-1.pdf, p41.

Please complete this form in full and return to protectingchildren@ofcom.org.uk.