Consultation response form

Your response

Volume 4: What should services do to mitigate the risk of online harms

Our approach to the Illegal content Codes of Practice

Question 12:	
i)	Do you have any comments on our overarching approach to developing our illegal content Codes of Practice?

Response: 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.[1] In the course of the OSA's passage 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 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

[1] 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

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 people 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 ask as private online law enforcement bodies, tasked with policing the speech of millions, far beyond pre-existing legal boundaries, which in our view will lead to a wave of privatised monitoring and censorship.

ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)

Response: No

Content moderation (User to User)

Question 18:

i)	Do you agree with our proposals?

Response: As we have set out above, we remain concerned by the impact the OSA, and by extension, Ofcom's proposals, will have on freedom of expression and privacy online, particularly in relation to the removal of 'user to user' content.

ii)	Please provide the underlying arguments and evidence that support your views.

Response:

As set out above, we have fundamental concerns with the requirement for social media platforms to make "illegal content judgements in relation to individual pieces of content for the express purpose of complying with the safety duties" (Vol. 4, 28). Many speech-related offences are highly dependent on context and intent which even the police and the courts find difficult to make determinations on (see *Chambers v Director of Public Prosecutions*, 2012). Social media platforms are not equipped to judge at scale (see our response to Question 49 for further detail).

We welcome Ofcom's proposal that content moderation requirements should only apply to illegal content "of which it is aware". In parts, the Online Safety Bill described platforms' need to "prevent" certain categories of content, 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"[2] 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[3]. Ofcom's approach, which states content moderation requirements should only apply to illegal content "of which it is aware" is an approach which is consistent with the legal standards previously held by liberal democracies when it comes to the regulation of online expression and in this regard, a positive step.

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

[3]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.open-rightsgroup.org/publications/legal-advice-on-prior-restraint-provisions-in-the-online-safety-bill/

We also welcome that Ofcom does not propose, at this stage, that companies should introduce measures to intercept or scan encrypted, private messages between users (Vol. 4, 29). End-to-end encryption is a vital safeguard for privacy and security and any there should not be any requirement on services to diminish this protection. Many private messaging services use end-to-end encryption to ensure that third parties (including the companies who operate the services and governments) cannot readily access users' private messages to one another. Suggestions that platforms might be required to break, erode or undermine the privacy and security provided to private messaging by end-to-end encryption are deeply troubling. This will create vulnerabilities within messaging services for criminals to exploit or could open the door to a greater level of surveillance.[4] International human rights bodies have recognised the importance of end-to-end encryption to protect the right to privacy and to promote the exercise of other rights. This is because being able to communicate safely and securely can be a precondition to being able to being able to communicate and express one's views – whether that is LGBTQ+ people seeking community in countries where homosexuality is illegal or journalists seeking to report on human rights abuses in places where there is limited press freedom.[5] The case law of the European Court of Human Rights (ECtHR) recognises the importance of anonymity in "promoting the free flow of ideas and information in an important manner" including by protecting people from reprisals for their exercise of freedom of expression.[6]

Interference with such companies' technical infrastructure is a matter of great legal and technical debate and would have a profound impact on rights. This does not mean users of such services are beyond the law – law enforcement agencies have a range of powers to seize devices, compel passwords and even covertly hack accounts and devices to circumvent end-to-end encryption.[7] End-to-end encryption means that the content of users' communications cannot be subjected to mass monitoring – and given the UK's commitment to upholding human rights and digital security, this should be protected.

We welcome Ofcom's decision to not mandate the use of automated tools for general content moderation (excluding the cases set out in Chapter 14, which we address further in the consultation), although we remain concerned that given the legal burden placed on user to user services to moderate content, many will inevitably have no choice but to use automated tools to fulfil their obligations. Ofcom appears to take a contradictory approach to acknowledging the impact automated tools will have on privacy, at one point noting the "important implications" they will have on privacy rights (Vol. 4, 20), but later stating that it "consider[s] that any interference with users' rights to privacy under Article 8 ECHR would be slight" (Vol. 4, 33). 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 as well as a wider chilling effect on user's speech (see our response to Question 20 for further detail).

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

[5]Written evidence submitted by Tech against Terrorism to the Joint Committee on the Draft Online Safety Bill, 14 December: https://committees.parliament.uk/publications/8206/documents/84092/default

[6]Delfi AS v Estonia [2015] EMLR 26, [147] and [149] quoted in legal opinion by Matthew Ryder KC and Aidan Wills on the human rights implications of client-side scanning, November 2022: https://www.indexoncensorship.org/wp-content/uploads/2022/11/Surveilled- Exposed-Index-on-Censorship-report-Nov-2022.pdf

[7]See Regulation of Investigatory Powers Act 2000, and Investigatory Powers Act 2016

We have a number of concerns about Ofcom's requirement that services prioritise flags from 'trusted flaggers' as "such complaints are likely to be accurate and to reflect the trusted flagger's assessment of harm" (45, vol. 4). 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 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.[8] 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 at 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 content in 'VIP' deletion lane and hasten censorship as a result.

Whilst we recognise that for the purposes of this consultation, trusted flaggers predominantly include law enforcement bodies and that a close relationship between these bodies and social media companies is important in the fight against crime online, these relationships must still be scrutinised closely to ensure human rights and civil liberties are protected. 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.[9] 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, undermining moderators' ability to make independent judgements. The reality of the power status of state authorities means that these flags are highly likely to result in enforcement action that suppresses speech. 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.

We are also concerned 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 to remove content at pace. Content removal decisions must be

[8]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

[9]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/

made cautiously and should be subject to appropriate scrutiny with detailed avenues of appeal available to those who have content removed. Any pressure on moderators to meet certain time goals will inevitably lead to rushed decisions. As well as the implications for freedom of expression, it is well documented that content moderators are already subject to serious workplace stress, trauma and pressure.[10] Requirements to speed up this process will likely exacerbate the problems with content moderation systems – both the toll on human moderators and the chilling effect on freedom of expression.

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 incentives 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, overbroad, and turns private companies into overzealous censors to avoid steep fines, leaving users with no judicial oversight or right to appeal."[11] 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."[12] 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 online", and argued that "censorship measures should not be delegated to private entities."[13] The law has also been criticised by the German broadcast media for turning controversial and censored voices into "opinion martyrs".[14]

iii) Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:No

[10]In Kenya labor dispute, workers who clean up toxic content on Facebook, TikTok and ChatGPT for $3 an hour go to court – Carlos Bajo Erro, El Pais, 5 August 2023: https://english.elpais.com/science-tech/2023-08-05/in-kenya-labor-dispute-workers-who-clean-up-toxic-content-on-facebook-tiktok-and-chatgpt-for-3-an-hour-go-to-court.html; Facebook moderator: 'Every day was a nightmare'- BBC News, 12 May 2021: https://www.bbc.co.uk/news/technology-57088382

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

[12]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

[13]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

[14]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

Content moderation (Search)

Question 19:
i) Do you agree with our proposals?
Response: We remain concerned by the impact 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. Many of the concerns we have set out in our response to Question 18 apply to this section and search services including the 'reliance on trusted flaggers'. As we have set out in our response to Vol. 5 of this consultation, designating content as illegal will be extremely challenging in some circumstances for many service providers. The requirement to undertake these assessments at scale will likely to lead to swathes of lawful content being erroneously downranked by search engines. We welcome Ofcom's decision not to recommend 'blanket deindexing' and acknowledgement that this would not be proportionate. However, we remain concerned that downranking content will still have significant impact on access to information, an important part of the public's right to freedom of expression and information.
ii) Please provide the underlying arguments and evidence that support your views.
Response: N/A
iii) Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: No

Automated content moderation (User to User)

Question 20:
i) Do you agree with our proposals?
Response: 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. Ofcom's decision not to recommend platforms use automated hash-matching systems for private communications channels at this stage is welcome. This could involve 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 surveillance through use of this technology. [15] It is vital that terrorism and CSEA content are removed from the internet. However, tackling such content 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 targetted and based on suspicion in line with the principles generally adhered to in liberal democracies. In a legal opinion commissioned by the free expression organisation, Index on Censorship, Matthew Ryder KC and

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

Aidan Wills of Matrix Chambers found that mandating these 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]).[16] The surveillance of millions of lawful users of private messaging apps has been found to require an extremely high threshold of legal justification, which content moderation purposes would be highly likely 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.

We have concerns that the requirement for platforms to use automated "fraud keyword detection" is disproportionate and is likely to result in the removal of lawful content.

We are additionally concerned that these tools will result in significant privacy intrusion for all users and are prone to errors and given the potential human rights risks, this will be an area for Ofcom to monitor closely.

ii)	Please provide the underlying arguments and evidence that support your views.

Response:

iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)

Response: No

Question 21:

i)	Do you have any comments on the draft guidance set out in Annex 9 regarding whether content is communicated 'publicly' or 'privately'?

Response: We are concerned that the draft guidance setting out whether content is communicated 'publicly' or 'privately' could impact on messaging services or functions, and could require services to scan users' messages.

Annex 9 asks platforms to make their own designation about whether content has been communicated publicly or privately, taking into account both statutory factors (outlined in Section 232(2) of the OSA) and "any other factors that Ofcom considers relevant" (Annex 9, 5). We are concerned that Ofcom's guidance makes no distinction between large 'group chats', such as those facilitated by messaging services such as WhatsApp, which are protected by end-to-end encryption and large open discussion forums. If large group messages are deemed 'public', services will be required under Ofcom's proposals to use automated technology to scan content that is end-to-end encrypted. As outlined in our responses to Q18 and Q20, such proposals would compromise the technical infrastructure that services such as WhatsApp use to keep users' messages safe and undermine the vital privacy protection offered by this technology to users.

The possibility that 'group chats' will be considered public content is made more likely by Ofcom's suggestion that: "The fact that content has not in fact been forwarded or shared with users of the

[16]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

service other than those who originally encounter it (or users of another internet service) does not mean that that content may not be shared or forwarded in such a way with ease." Ofcom's suggestion that content could be considered public by the possibility, rather than actuality, of privately communicated content being shared is concerning.

The guidance also states that privately communicated content could later be considered to be communicated publicly, as any designation "may change over time" (Annex 9, 5). This gives users very little certainty over the privacy of their communications, and what level of intrusion it will be subject to from platforms. It is also difficult to envision how such proposals would be compatible with platforms that use of end-to-end encryption. These proposals should be redrafted to ensure that large 'group chats' are not considered publicly communicated content in order to safeguard the key privacy protections afforded by end-to-end encryption.

ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)

Response: No

Do you have any relevant evidence on:

Question 26:	
i)	An effective application of hash matching and/or URL detection for terrorism content, including how such measures could address concerns around 'context' and freedom of expression, and any information you have on the costs and efficacy of applying hash matching and URL detection for terrorism content to a range of services.

Response: We welcome Ofcom's decision not to mandate the use of automated content-moderation technology for the purposes of combatting terrorism content given the likelihood of over-removal and the inherent threat to freedom of expression.

ii)	Please provide the underlying arguments and evidence that support your views.

Response: As we will highlight in our response to questions relating to volume 5, in our view, social media platforms are ill-equipped to make determinations on the legality of speech, particularly when it comes to making judgements on when expression may or may not fall foul of speech-related criminal offences. Under the threat of penalties, it is likely that these companies will over-moderate and censor entirely lawful expression out of an abundance of caution.

Schedule 5 of the Online Safety Act sets out the suspected terrorism offences that platforms must take down on their sites. They include section 12(1A) of the Terrorism Act 2000 which makes it a criminal offence to express an opinion or belief supportive of a proscribed organisation) and 13(1A) of the Terrorism Act 2000, which makes it a criminal offence to publish an image of the uniform of proscribed organisation. These are complicated offences which law enforcement bodies and courts must make careful judgements on, balanced against their obligations set out in human rights law. They are not offences which automated content-moderation systems can definitively identify. It is vital that terrorism content is removed from the internet. However we welcome Ofcom's decision not to mandate platforms to use automated content-moderation technology to detect and remove material of this nature. Genuine terrorist material online, which constitutes a security threat to the public, should not be only dealt with by companies in Silicon Valley but the police and other security bodies. Offences of this nature should then be determined by the full rigour of the criminal justice system and cannot accurately be determined

by automated content-moderation technology in a way which will not have a negative bearing on freedom of expression.
iii) Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: No

User reporting and complaints (U2U and search)

Question 28:
i) Do you agree with our proposals?
Response: We believe Ofcom should go further in setting minimum standards for appeals and complaints processes.
ii) Please provide the underlying arguments and evidence that support your views.
Response:

Whilst it is welcome that the Online Safety Act compels online intermediaries to offer an appeals process to users who have had content restricted, in practice this is already the case on most major social media platforms. However, these processes are often ineffectual, automated, lack clear process and content is rarely assessed in the full context in which it was posted.

We believe that Ofcom should go further in creating minimum standards for platforms appeals processes in line with its duties to uphold and promote freedom of expression.

The Santa Clara Principles (2021), drafted by human rights organisations and academics establish some basic principles for centralised content moderation systems, to ensure they are compliant with human rights standards. The principles state the importance of appeals processes in protecting freedom of expression. They state that user notice to those who have contravened a platform's rules should include the following:

- URL, content excerpt, and/or other information sufficient to allow identification of the content actioned.
- The specific clause of the guidelines that the content was found to violate.
- How the content was detected and removed (flagged by other users, trusted flaggers, automated detection, or external legal or other complaints).
- Specific information about the involvement of a state actor in flagging or ordering actioning. Content flagged by state actors should be identified as such, and the specific state actor identified, unless prohibited by law. Where the content is alleged to be in violation of local law, as opposed to the company's rules or policies, the users should be informed of the relevant provision of local law.

The Santa Clara Principles also state that appeals processes should incorporate the following:

• A process that is clear and easily accessible to users, with details of the time-line provided to those using them, and the ability to track their progress.

• Human review by a person or panel of persons who were not involved in the initial decision.

• The person or panel of persons participating in the review being familiar with the language and cultural context of content relevant to the appeal.

• An opportunity for users to present additional information in support of their appeal that will be considered in the review.

• Notification of the results of the review, and a statement of the reasoning sufficient to allow the user to understand the decision. Only through intermediaries following due process and applying a rules-based approach to content moderation can users rights be fully respected online.[17]

iii) Is this response confidential? (if yes, please specify which part(s) are confidential)

Response:No

Terms of service and Publicly Available Statements

Question 29:

i) Do you agree with our proposals?

Response: We welcome Ofcom's approach in ensuring that rules platforms use to moderate content on their sites are accessible and transparent, however we have some freedom of expression concerns which engage this section.

ii) Please provide the underlying arguments and evidence that support your views.

Response:

The terms of service model, used by many large social media platforms to govern their sites, does not lend itself to accessibility or clarity when it comes to content moderation decisions. Whilst some platforms try to present their terms through more accessible "community guidelines", many do not or simply locate these pages where they are obscured from users' view. Often changes to platforms' rules are also published in places where users are unlikely to see them.

When it comes to the permissibility of speech online, major internet intermediaries need digital constitutions that reflect the foundational values of the democracies they serve. This means content policies should reflect human rights principles and avoid limiting expression beyond the limitations of the law. These constitutions should clearly presented to users upon first access to the site, made accessible to users and should be referred to in all content moderation decisions.

Currently, the terms of service model effectively gives most platforms absolute power and complete discretion as to their application of it. This needs to change. We believe that major internet platforms should adopt rule of law principles for enforcement. Ofcom should endeavour to promote rule enforcement that centres transparency of rules, foreseeability of their

[17]The Santa Clara Principles, 2021, https://santaclaraprinciples.org/

application, fairness of processes, the right to appeal, and equal and consistent application of the rules.

In particular, when setting out their rules, platforms should make the text easy to understand. Rules should be clearly defined and refrain from being subjective. Users should be actively notified by the platform as to any rule changes.

By ensuring that rule of law principles are embedded in platforms' processes, in a way which is clear to users, fundamental rights can be protected online.

iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)

Response:No

Volume 5: How to judge whether content is illegal or not?

The Illegal Content Judgements Guidance (ICJG)

Question 49:

i)	Do you agree with our proposals, including the detail of the drafting?

Response: Big Brother Watch remains deeply concerned that requirements for online platforms to judge what constitutes illegal content will result in mass surveillance and the censorship of lawful content due to platforms inability to make determinations on the permissibility of speech in this way.

ii)	What are the underlying arguments and evidence that inform your view?

Response: Removing so called "illegal content" for the purposes of complying with the regulatory system covers not only content which reaches conviction in a criminal court, but anything that a platform determines could be illegal. This system, which constitutes the state requiring private companies to make determinations on what constitutes illegality undermines the rule of law and poses serious threats to freedom of expression. Whilst the identification of illegal material may be clear and obvious in some cases, in many others defining communications of this nature is a complex matter traditionally reserved for law enforcement bodies and the judicial system.

Services must have "reasonable grounds to infer" that content is illegal before removing it. The Ofcom consultation acknowledges that this is a "new legal threshold" that goes further than the standard set out in UK courts of being "beyond reasonable doubt". This is significantly below the ordinary standard of proof required to determine that that a crime has been committed. Under this definition, platforms will inevitably censor entirely lawful speech. Social media companies, and individual content moderators do not have the competency or authority to make determinations of this kind. The consultation notes "it is often hard to establish" whether the tests set out in Section 192(5) of the OSA to make this designation are met (Vol. 5, 7). It is for this reason that such decisions have always been taken by the courts. Outsourcing this decisions to private companies is wholly inappropriate, regardless of any guidance provided by Ofcom.

The obligation for platforms to determine what constitutes illegality will become problematic around the limitations of free expression. Offences set out in the Public Order Act (1986) criminalise those who "stir up hatred" through their use of "words, behaviour or written material"

These offences have been carefully developed through multiple rounds of rigorous Parliamentary scrutiny in order to protect minority groups. The full rigour of the criminal justice system and referral to established case law are necessary to make a conviction under offences of this nature.

Another example, the Communications Act (2003), criminalises communications that are deemed to be "grossly offensive". This legislation has proved to be deeply controversial since it was commenced and has resulted in the criminalisation of speech that causes serious offence. In the case of the well-documented "Twitter joke trial", a man was prosecuted after learning that an airport from which he was due to travel was closed due to snow-fall and joking that he would "blow the airport sky high". In Chambers v Director of Public Prosecutions (2012), the High Court overruled the verdict of a magistrate's court that had found the defendant guilty of sending a "menacing electronic communication" under the Communications Act.[18] This demonstrates the complexity of the law in this area and the care that is required when considering the permissibility of speech.

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, police racism, jokes about gender stereotypes, sexuality, and statistics about crime have all been flagged by platforms as inciting hatred and wrongly removed.[19]

Further, the courts, Crown Prosecution Service (CPS) and the police are all bound by a duty under the Human Rights Act 1998 to act in accordance with the European Convention on Human Rights, including protecting the right to freedom of expression. No equivalent duty falls upon the platforms.

The risks to free expression are clear. Under rigorous obligations to protect people from illegal content on their sites, online intermediaries, who are not qualified to establish what constitutes illegal speech will over-remove content on their platforms under the threat of penalties. The consequential impact on free speech will be profound.

iii) Is this response confidential? (if yes, please specify which part(s) are confidential)

Response: No

[18]Robin Hood Airport tweet bomb joke man wins case - BBC News, 27 July 2012, https://www.bbc.co.uk/news/uk-england-19009344

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