CIVIL SOCIETY COMMITTEE STAGE BRIEFING ON THE ONLINE SAFETY BILL FOR HOUSE OF LORDS: ILLEGAL CONTENT SAFETY DUTIES AND PRIOR RESTRAINT – Supported by Wikimedia UK, Electronic Frontier Foundation, Index on Censorship, and Open Rights Group.

Published by Open Rights Group - Open Rights is a non-profit company limited by Guarantee, registered in England and Wales no. 05581537. The Society of Authors, 24 Bedford Row, London, WC1R 4EH. (CC BY-SA 3.0)

Author Dr Monica Horten

For Further information please contact Dr Monica Horten on 07787 512 887 – monica@openrightsgroup.org

18 April 2023

Free speech is a precious right that is core to our democracy. In these times when the Internet and social media are important platforms for speech, it is vital to ensure that any actions addressing harms do not also harm our free speech rights. Finding the balance between freedom of expression, and other rights such as privacy, security and children's rights, is crucial.

When the words one says, one's personal photos and videos, are screened for compliance with state-mandated criteria, these are the beginnings of interference with free speech rights online. It may result in a decision to remove or restrict the content. Where screening is done by algorithms, the decision will be taken without context on the basis of a binary, non-contextual determination.

Algorithmic screening of content is built into the Bill, and whilst it is not an explicit requirement, the assumption is that it will be needed for compliance purposes. It is the intended interpretation of the term " proactive technologies". Any proactive screening of content would need to be continuous, all day, every day – a form of general monitoring. All of this raises concerns regarding interference with free speech rights.

Social media providers will use algorithmic screening on all priority and primary priority illegal content as detailed in Schedules 5,6, and 7. The broad and vague criteria for determining illegal content , and the requirement to assess the mental aspects of offences, could increase the number of data points collected about the user and their activities, resulting in a greater intrusion into their privacy. The lack of requirement for evidenced decisions would lead to arbitrary removal of lawful speech.

Our particular concern relates to a requirement to "prevent" users seeing or engaging with illegal content. In this context, the algorithmic screening would be carried out prior to the content being uploaded. Put simply, as the user begins to upload their content, the system would sweep in and check it out, and potentially remove it before it has been communicated. This is a modern, 21st century form of prior restraint.

Prior restraint is a particularly severe restriction on free speech. It is an action that prohibits speech, or any other form of expression, before it can take place[1]. With regard to online speech, it means banning speech, text, images, or videos, before being published or posted.

The individuals who posted it may never have a possibility to appeal, because, despite the Bill allowing for them to so, there is no requirement for them to be informed about the restriction. This is incompatible with human rights law which requires that procedural safeguards are in place to prevent arbitrary interference with freedom of expression[2]. The overall outcome would be a chilling effect on public debate.

Our proposed amendments will protect against prior restraint of lawful speech by tightening the evidential requirements and will ensure that there are procedural safeguards in place that will enable users to challenge decisions through complaint or appeal.

AMENDMENTS TO SAFEGUARD AGAINST PRIOR RESTRAINT

We urge support for the following amendments:

Lord Stevenson of Balmacara, Lord Clement-Jones and Baroness Stowell
After Clause 184
No obligation to undertake general monitoring *Nothing in this Act introduces an obligation on a regulated service to undertake general monitoring of content on its service." Member's explanatory statement: This amendment is to probe whether social media platforms and other regulated services will be required to undertake general monitoring of the activity of their users.*

Lord Moylan
Clause 9
Page 7, line 30, *leave out "prevent individuals from" and insert "protect individuals from harms arising due to them"*
Member's explanatory statement: This amendment, along with the other amendment to Clause 9 in the name of Lord Moylan, adds a requirement to protect individuals from harm, rather than monitoring, prior restraint and/or denial of access. Further obligations to mitigate and manage harm, including to remove unlawful content that is signalled to the service provider, are unchanged by this amendment.

1 Council of Europe Factsheet May 2018 'Prior restraints' and freedom of expression: the necessity of embedding procedural safeguards in domestic systems' https://rm.coe.int/factsheet-on-prior-restraints-july2018-docx/16808c1691

2 See Footnote 1

Page 7, line 40, *leave out paragraphs (a) and (b) and insert "take down illegal content, swiftly after the provider is specifically alerted to the presence of that content and its illegality, or becomes aware of it in any other way".*
Member's explanatory statement : This amendment, along with the other amendment to Clause 9 in the name of Lord Moylan, adds a requirement to protect individuals from harm, rather than monitoring, prior restraint and/or denial of access. Further obligations to mitigate and manage harm, including to remove unlawful content that is signalled to the service provider, are unchanged by this amendment.

We recommend the Bill is further amended as follows:

In Clause 170 (5) and (6)
Providers judgements about the status of content
(5) In making such judgements, the approach to be followed is whether a provider has **sufficient evidence** ~~reasonable grounds to infer~~ that content is content of the kind in question (and a provider must treat content as content of the kind in question if reasonable grounds sufficient evidence for that inference exist).
(6) ~~Reasonable grounds for that inference~~ *Sufficient evidence* exists in relation to content and an offence if, following the approach in subsection (2), a provider—
(a) has ~~reasonable grounds to infer~~ **sufficient evidence** that all elements necessary for the commission of the offence, including mental elements, are present or satisfied, and
(b) does not have ~~reasonable grounds to infer~~ **sufficient evidence** that a defence to the offence may be successfully relied upon.
Reason: The amendment is to avoid arbitrary removals or restrictions of lawful content by introducing a requirement for "sufficient evidence".

In Clause 17 (2)
add (d) "*provides for a notification to any user whose content has been removed or restricted, including the nature of the restriction, the length of time that the restriction will be in place, an evidenced justification and information on how they may appeal.*"
Reason: The amendment adds a procedural safeguard against arbitrary removals or restrictions of lawful content.

QUESTIONS OUTSTANDING
Clause 9 Safety Duties for illegal content raise significant implications for democratic discourse and freedom of expression, although the lack of clarity on the face of the Bill means it is not evident. For this reason, we ask that Peers put the following questions to the Secretary of State:

- Can the Secretary of State explain what would constitute "preventing users encountering" content [Clause 9 (2)(a)], and how providers will be expected to comply with this measure? Can the Secretary of State confirm if the expectation is that content will be scanned and removed on upload?

- Can the Secretary of State confirm whether the government intends to retain the prohibition on a general monitoring obligation in UK law?

- Can the Secretary of State clarify how immigration and public order offences (Priority illegal content, Schedule7) will be determined by providers?

- Under what circumstances would Ofcom require the use of "proactive technology" (content moderation systems that seek out and identify content, using artificial intelligence and algorithms)? [Clause 202(1), (2), and (10)].

- Can the Secretary of State clarify what procedural safeguards exist for users whose post or content was unlawfully removed on upload and how they could seek redress, including judicial redress? [Clause 17(4)]

HOW THE BILL REQUIRES CONTENT SCREENING PRIOR TO PUBLICATION

1. The Online Safety Bill will require providers to "*prevent users encountering*" priority illegal content [Clause 9(2)(a) Safety Duties about illegal content]. This requirement is distinct from the obligation to "*swiftly take down*" content when it has been reported or to proactively search for it and take it down in order to *"minimise the length of time*" on the platform.

2."Encounter" means *"read, hear, view or otherwise experience*" the content. [Clause 163(5)] There is no explanation on the face of the Bill as to what should be done by the provider, but the language does suggests that the platform has to stop users from seeing, hearing or experiencing the content at all. The way to do that is to stop it appearing on the platform in the first place[3].

3. In order to comply, social media companies will need to screen content prior to it being posted onto the platform. This would be done by algorithmic screening of the content whilst it is being uploaded. The content would be intercepted on upload, checked and where deemed illegal, removed. It could then be reported to the National Crime Agency, where the Bill requires this to happen.

4. The systems that do this are known as *content moderation systems*. They are defined in the Bill under *"proactive technology"* (Clause 202(2)]). The definition in the Bill states that they use artificial intelligence and algorithmic programming techniques. Content moderation systems are designed to proactively seek out and identify the proscribed content, and to take action against the content such as remove it or report it. A content moderation system that screens content whilst uploading, is sometimes referred to as an "upload filter".

5. Content moderation systems check the content to see if it is illegal, as defined in Clause 53. Priority illegal content (Clause 53 (10) is detailed in Schedules 5,6, and 7. These systems can only take binary decisions, and are incapable of deliberating nuanced context. One technique known as "perceptual hashing" [4]is used to match the images against a database of "hashes" or digital fingerprints of images that have been pre-determined to meet the

3 It could also be done by age verification, as mandated in Clause 11.
4 Wikipedia: Perceptual hashing https://en.wikipedia.org/wiki/Perceptual_hashing

criteria for the content to be removed. Where the system finds a match, it will remove the videos or images, or forward them to a human moderator for further checking. These images will not appear either on the user's own account or in other users' timelines.

6. A significant risk is that there is inherent imprecision in systems that automatically block or remove content based on an algorithm. Or that more generally speaking, technical filter systems are prone to error, because no algorithm can perform the kind of contextual and legal analysis required to distinguish lawful from unlawful uses.

7. In order to catch the illegal posts, continual monitoring is required. This means all of the users and the content they post, all of the time, 24 hours a day, 365 days a year. This is known in legal terms as "general monitoring". It is considered excessive and a risk to free speech. [5]

8 The law does allow providers to monitor their own platforms for specific purposes, but the position is different if the State is requiring them to conduct general monitoring. To date, UK law has prohibited the State from imposing an obligation for such general monitoring on platform providers. This came into force under the UK transposition of the EU E-commerce Directive[6]. It formed a key foundation stone in Internet law, aimed at safeguarding users' freedom to expression and their right to information. It plays an essential function in preventing the emergence of checkpoints that would serve to entrench State or corporate power over online speech. Following the UK's withdrawal from the EU, and the proposed repeal of EU law, this prohibition on a general monitoring obligation will fall away. The UK government had committed to retaining it, but currently its position is not clear.[7]

9. The whole process will be underpinned by a code of practice written by Ofcom. A platform that implements the measures in this code of practice is deemed to have complied with the Bill. Ofcom may include in its Code of Practice a requirement to deploy these systems (Schedule 4, 12 *Proactive Technologies*).

10.Ofcom may mandate the use of proactive content moderation systems in its Codes of Practice [Schedule 4(12)]. The regulator has powers to impose *"accredited technology"* (Clause 110) which is a content moderation system designed to government standards (Clause 202 (11)) for the purpose of seeking out terrorism content or CSEA material on public platforms.

11. There are risks to Ofcom's regulatory independence[8] by the broad powers granted to the Secretary of State. This raises serious questions around political control over a process to

---

5 Christina Angelopoulos and Martin Senftleben (202) The Odyssey of the Prohibition on General Monitoring Obligations https://www.cipil.law.cam.ac.uk/press/news/2020/10/odyssey-prohibition-general-monitoring-obligations-between-article-15-e-commerce and Herbert Zech 2021 General and specific monitoring obligations in the Digital Services Act https://verfassungsblog.de/power-dsa-dma-07

6 Article 15 of the E-Commerce Directive states that the government may "not impose a general obligation on providers to monitor the information which they transmit or store, nor a general obligation to seek facts of circumstances indicating illegal activity".

7 Graham Smith (2017) Time to speak up for Article 15 https://www.cyberleagle.com/2017/05/time-to-speak-up-for-article-15.html and Graham Smith (2021) Corrosion-proofing the UK's intermediary liability protections https://www.cyberleagle.com/2021/02/corrosion-proofing-uks-intermediary.html

8 Letter from the Rt Hon Baroness Stowell of Beeston, Chair of the Communications and Digital Committee, to Rt Hon Michelle Donelan, Secretary of State for Digital, Culture, Media and Sport, 30 January 2023

determine illegal speech and censor it. The recent statement by the Secretary of State about banning images of "small boats" crossing the Channel carrying asylum seekers, is illuminating, as discussed below. [9]

INCENTIVES TO BLOCK BEFORE PUBLICATION

12. Illegal content is defined by a list of priority offences in Schedule 5 (Terrorism), Schedule 6 (CSEA) and Schedule 7 which lists 33 criminal offences that include assisting illegal immigration, public order, and national security as well as harassment, assisting suicide, and threats to kill.

13.The Joint Committee on Human Rights[10] identified the difficulties in determining illegality for the purposes of the Bill, asking how a provider of user-to-user services would identify an offence under Section 5 of the Public Order Act 1986 in a social media post? Similar questions apply to all other offences in Schedule 7.

14. The House of Lords must support amendments that will prevent illiberal legislation passing unchecked. The content is illegal if the use of "words, images, speech or sounds" amounts to a relevant offence, or if the possession or dissemination of that content amounts to an offence.

15. A determination of illegality will be made by the companies themselves, who will then take action to remove the content. They must have "reasonable grounds to infer" [170(5)] that the content meets the illegality threshold. If it does, the content must be removed. This was confirmed by the Secretary of State in a public statement on 17 January.[11]

16. "Reasonable grounds" includes the "mental element" or criminal intent, sometimes referred to by the Latin *"mens rea*". There is no requirement to examine evidence. Instead, determinations of illegal content "are to be made on the basis of all relevant information that is reasonably available to a provider."

17. It goes without saying that this is a far lower threshold of illegality than required in a court, where decisions are deliberated on the basis of evidence. The complex and nuanced decision-making around the "mental element" (Clause 170) or the criminal intent, will be a challenge for proactive content moderation systems.

18. This would entail a far lower threshold than a court, and a more limited factual basis for a decision[12]. A court would make its judgment by deliberation and reference to established facts, whereas a content moderation system makes a binary 'yes' or 'no' decision. It makes for a flawed decision-making process, especially given the possibility that the content is removed, with little possibility for the user to challenge it.

---

9 Online Safety Update 17 January 2023, Michelle Donelan, https://questions-statements.parliament.uk/written-statements/detail/2023-01-17/hcws500

10 Letter from Rt Hon Harriet Harman MP, Chair of the Joint Committee on Human Rights, to Rt Hon Nadine Dorries, Secretary of State, for Digital, Culture, Media and Sport DCMS, 19 May 2022

11Written Statement, Michelle Donelan, 17 January 2023 https://questions-statements.parliament.uk/written-statements/detail/2023-01-17/hcws500

12 Graham Smith, https://www.cyberleagle.com/2023/01/positive-light-or-fog-in-channel.html

19. The Bill asks providers to undertake a risk assessment[13] for every one of the 33 offences listed in Schedule 7, as well as the multiple terrorism offences and child sexual abuse material (CSEA) and then implement the outcome of the risk assessment in their systems. For small providers, this is a high burden, as their AI would need to be trained to identify content related to each specific offence. Moreover, the latest content moderation systems are being designed to offer considerable flexibility for platform providers to set their own standards. Whilst this may sound reasonable, it risks a patchwork of criteria resulting in considerable uncertainty for lawful users who post on multiple platforms.

20. There are strong incentives for providers to comply. Failure to do so risks heavy fines (£18 million or 10 per cent of qualifying worldwide revenue), or imprisonment of their management[14]. These tough sanctions will incentivise over-moderation of content – taking down content where there is uncertainty over its illegality or is picked up by the algorithms as a 'false flag".

21. The increased number of data points that would have to be collected in order for the determination of illegality to be made, indicates an increasing risk of surveillance and privacy intrusion.

22. Private actors should not be taking these decisions. There are serious concerns that providers will take an overly-cautious approach to content removal where they have difficulty in determining illegality, acting under pressure of large fines and criminal liability. The combination of over broad definitions, proactive technology application and penalties for non-compliance is likely to result in a form of prior restraint.

THE CHILLING EFFECT OF PRIOR RESTRAINT

23.The proposed government amendment to address illegal immigration and ban images of small boats of asylum seekers crossing the Channel[15], is one example that illustrates the way that content could be overly-cautiously moderated.

24.The government said in a statement on 17 January,[16] that an amendment to the Bill would result in the removal of images and videos of small boats of asylum seekers crossing the Channel, that show the activity "in a positive light".

25.It is being drafted in response to a previous amendment proposed by the MP for Dover, Nathalie Elphicke, that called for the removal of "content that may result in serious harm or death to a child while crossing the English Channel with the aim of entering the United Kingdom in a vessel unsuited or unsafe for those purposes." It was positioned within Clause 11 "*Safety duties protecting children*".

---

13 Clause 8(5)

14 Jacqueline Rowe, Global Partners Digital: Policy Briefing - The proposal to expand criminal liability for social media managers in the Online Safety Bill

15 Censorship fears over plan to keep Channel people-smugglers off social media https://www.theguardian.com/media/2023/jan/18/banning-channel-tiktok-traffickers-risks-censorship-uk-campaigners-say

16Written Statement, Michelle Donelan, 17 January 2023 https://questions-statements.parliament.uk/written-statements/detail/2023-01-17/hcws500

26.However, the proposed government amendment says nothing about boats. It will insert Section 24 of the 1971 Immigration Act. It concerns unlawful entry to the UK, and has been updated by the Nationality and Borders Act 2022 Section 40, making it a criminal offence to arrive in the UK without a valid entry clearance (and adds to an existing provision in Schedule 7(22) headed "Assisting Illegal Immigration", Section 25 of the 1971 Act). The government's rationale is that images of small boats would be considered "inchoate offences" such as aiding, abetting, and encouraging (Schedule 7 (33)). The government amendment would also apply to videos of people trying to enter the UK by climbing aboard lorries.

27.There are many legal question marks around this approach, and how illegality would be determined. Legal opinion suggests that it is not clear how a provider would determine whether or not the image shows the activity "in a positive light", or whether a video is capable of encouraging an offence of this nature[17].

28.Other Schedule 7 offences follow a similar logic. For example, how should one interpret in this context, the Public Order Act 1986, Section 5, which is also listed in Schedule 7? This point has already been raised by the Joint Committee on Human Rights. [18] The Public Order Act 1896, Section 5, relates to public protests.

29.There are further issues around interpretation when it comes to Schedule 5 Terrorism offences. For example, determining whether someone using a rifle is a terrorist, or a soldier, would require context[19]. There are many instances where video and images of conflict zones has been removed by online platforms on the basis of their terrorism community standards, when in fact the content is lawful and it is in the public interest that it stays online. For example, this happened in the Syrian conflict. Providers who are operating under threat of imprisonment or very large fines for failure to remove, are likely to over-moderate, and where they are blocking on upload, there is a real risk of public interest content being suppressed.

30.Importantly, if the images are determined to be illegal, then they would be blocked on upload, prior to publication, under Clause 9(2), as outlined above. This is a form of prior restraint. It highlights the likely chilling effect on public debate if such images were banned prior to publication. Prior restraint will work together with other measures in the Bill, such as age verification, to restrict access to content and raises deep concerns not only for freedom of expression, but for British democracy.

SAFEGUARDS FOR FREE SPEECH IN THE CONTEXT OF PRIOR RESTRAINT

31. The possibility of prior censorship raises serious questions around users' freedom of expression. In 2022, people exercise their free speech rights online and on social media

17 Graham Smith, (2023) Positive light or fog in the Channel https://www.cyberleagle.com/2023/01/positive-light-or-fog-in-channel.html

18See also Letter from Rt Hon Harriet Harman MP, Chair of the Joint Committee on Human Rights, to Rt Hon Nadine Dorries, Secretary of State, for Digital, Culture, Media and Sport DCMS, 19 May 2022

19 See Independent Terrorism Reviewer (2022) Missing Pieces: Terrorism Legislation and the Online Safety Bill https://terrorismlegislationreviewer.independent.gov.uk/missing-pieces-terrorism-legislation-and-the-online-safety-bill/

platforms or user-to-user services[20]. It is the duty of the State to safeguard those rights, and protect individuals against interference with their rights. At the same time, there are rules providing for situations where the State may have legitimate reason to restrict speech rights. Prior restraint - banning by algorithmic screening before publication- is a serious interference with free speech rights. The rules should say clearly and precisely why the content is to be banned, and provide for procedural safeguards to address cases where lawful content has been restricted.

32. The Bill needs strengthening on both counts. As it stands, it does not recognise users' positive rights to free speech, or that interference with free speech rights could occur. It only requires that a complaints procedure is established but the way it will operate is left to Ofcom to determine in a Code of Practice.

33. The Bill does state that users can complain about the removal of their content. Clause 17(4) (c) allows them to make a complaint if the basis of the take-down was "illegal content". Clause 17(4)(e) allows a complaint for content removals using proactive technology. However, in the case where the content was removed prior to upload, how would they be able to challenge it, when they may not even know what happened? We recommend that Clause 17 is amended to provide for a notification to users whose content has been restricted, including rights of appeal.

34. The Bill would be considerably strengthened with a provision for users to be notified about the decision to remove their content. This would align it with international human rights standards. A notification should include a clear and specific statement of reasons including grounds for illegality with evidence. Their right to appeal it. and to judicial redress, should be confirmed on the face of the Bill. [21]

20 Index on Censorship: A Legal Analysis of the Impact of the Online Safety Bill on Freedom of Expression https://www.matrixlaw.co.uk/wp-content/uploads/2022/05/Legal-analysis-of-the-impact-of-the-Online-Safety-Bill-on-freedom-of-expression.pdf See also Council of Europe Recommendation CM/Rec (2014) 6 of the Committee of Ministers to member States on a Guide to human rights for Internet https://www.coe.int/en/web/freedom-expression/guide-to-human-rights-for-internet-users

21 Council of Europe Recommendation on Internet Freedom [CM/Rec(2016)5] .

STAGE	ONLINE SAFETY BILL	REAL WORLD
Investigation	• Social media companies will use content moderation systems ["proactive technology" Clause 202(2)] to seek out and identify illegal content using artificial intelligence • continual proactive monitoring of users is required to achieve compliance meaning all content and users, would have to be monitored all of the time in order to flag illegal posts	• Victims report allegation of a crime to police (or another relevant authority) • Police can arrest and interview a suspect if they have reasonable suspicion that the individual has committed the offence linked to the allegation • The suspect has an opportunity to explain their version of events or indeed decline to comment
Adjudication	• Content will be intercepted by the algorithm on upload by the user • Content may be checked by a moderator (with no requirement they are legally-trained or based in the UK) • the moderator deems the content illegal if they can "reasonably infer" illegality from the available information	• Following a policing investigation based on statutory powers, a decision is made (mostly in conjunction with the Crown Prosecution Service) whether (a) there is sufficient evidence of the crime (b) it is in the public interest to charge the suspect • There is a requirement that the case can be proven 'beyond reasonable doubt'
Action	• Content moderation systems take binary decisions, and are incapable of nuanced decision-making around the "mental element" (Clause 170) of offences	• A criminal case proceeds to the relevant court where a tribunal of fact and a tribunal of law decide whether a criminal offence has been proven • The defendant in the proceedings benefits of being assumed to be innocent before proven guilty • The defendant has the opportunity to present their defence and or explanation in response to the allegation
Outcome	• With no legal requirement of (i) reasonable suspicion, (ii) evidence or to (iii) investigate and with the threat of large fines and criminal liability of executives, swathes of lawful content will also be removed • There is no requirement to address the individual offending behaviour nor assess the potential parallel danger to the intended victim of the content.	• If the case is proven beyond reasonable doubt, the convicted individual faces sentencing and any relevant corroborative restorative justice measures to prevent reoffending. • The victim is provided the opportunity to express the impact on them and on the punishment to the offender

[Table credit: Index on Censorship].