ONLINE SAFETY BILL ROUNDTABLE

Prior Restraint

May/August 2023

On 24 May 2023, Open Rights Group (ORG) held an online roundtable to discuss proposals for the Online Safety Bill that ORG argues will amount to prior restraint. The Bill proposes that content deemed illegal under the legislation should be prevented from appearing on the platform, thus controlling expression and a form of censorship.

Participants ranged from across the spectrum of civil society, including representatives from groups working on over-policing, youth justice, counter-terrorism, and women's and migrants' rights.

The event was chaired by Dr Monica Horten, Policy Manager (Freedom of Expression), who also briefed participants on the unintended consequences of prior restraint and the mechanism for removing illegal content. Meg Foulkes, Head of Policy and Litigation, also discussed the amendments addressing self-harm.

A discussion took place around the proposals and their implications.

Prior Restraint in the Bill

1. Prior restraint is the action of banning content before publication. In the online world, that can refer to filtering or screening content whilst it is being uploaded to a platform or service. The key clause in the Bill is section 9(2a) . It asks platforms to take or use proportionate measures relating to the design or operation of the service to *prevent individuals from encountering* priority illegal content by means of the service.

2. The Bill suggests proactive systems will seek out and detect the content that they want to remove;. This content will be intercepted and blocked or taken down before it gets onto the platform. It's a particular form of content moderation that would be carried out via automated, algorithmic/AI-driven systems.

3. The relevant clause in the Bill is clause 9(2), describing the safety duties for illegal content, which is detailed in schedules 5, 6 and 7; which specify the offences against which content will be assessed for take down (based on

corresponding legislation addressing those offences “offline”); and communications offences in the main body of the Bill that includes a new amendment on self harm. Section 193 sets out how providers are to make determinations of illegality.

4. AI-driven/algorithmic systems make mistakes because these systems are blind to context and intention and “false flags” are common. To demonstrate the difficulties in the mechanism proposed, these systems can’t tell the difference between commentary or a radical threat to national security, for example.

5. Online platforms are likely to be overcautious as the Bill mandates harsh consequences for non-compliance – criminal liability and even jail time.

6. A closer look at the offences shows the ambiguities that may shut down public debate.

7. “Assisting illegal immigration” has been an offence in the bill since its inception but the introduction of a new amendment prompted the government to explain how it would require online platforms to take down images of small boats with asylum seekers crossing the Channel, as posting these images would be an inchoate offence.

8. Several offences from the The Public Order Act 1986 are in the bill but it is unclear what kind of content could be taken down as a result. For example, correlate these offences with the pre-protest arrests prior to the King’s Coronation, supposedly in line with the new Public Order Act, which was implemented within a week of its passing, although the police had little understanding of what they were supposed to be implementing. Would we see images of public protest being removed prior to publication and what would be the effect of that on public discourse?

9. In the context of the Online Safety Bill, platforms – private companies – will enforce the legislation with potentially little knowledge of what they’re to enforce. Lord Moylan tabled an amendment to remove the offences under section 5 of the Public Order Act, however it was not voted on.

10. Commonly, platforms search for symbols representing terrorist/hate groups e.g. pictures of swastikas from Nazi Germany, an ISIS flag. A case in 2016 saw a war photographer’s picture of a bomb-damaged village in Syria with an ISIS flag taken down.

11. The Facebook Oversight Board overturned a decision by the platform to take

down a post including a quote from Hitler propagandist Joseph Goebbels, being used to comment on the Trump administration and the US context. Similar posts at the time in the UK were also taken down, demonstrating the confusion that can occur when arbitrarily assessing content.

12. The Independent Reviewer of Terrorism has commented on lost context when trying to determine terrorism offences in the online sphere. For example, someone who is training somebody to shoot a rifle could be a member of rifle club and doing that legally. But the platform doesn't know because it doesn't know the context that rifle training is taking place.

13. The debate around Shamima Begum will be worth watching as it remains unclear if she is considered a terrorist for the purposes of this Bill, and therefore whether content discussing her case could be taken down.

Measures to address self harm content

14. Clause 185 is a communications offence to address self-harm content. It was highlighted in order to demonstrate the difficulties for online platforms determining illegality or whether an offence had been committed.

15. The amendment sets out the criminal law apparatus for identifying those accountable for the production of illegal content that results or could result in self harm using the Offences Against the Person Act 1861 and the threshold of grievous bodily harm (GBH) – section 18 and 20 are the two offences in that; distinguished by the requirement of specific intent for section 18 i.e. causing grievous bodily harm with intent.

16. Intent is subjective and potentially contentious; the offences against a person requires GBH is performed, unlawfully and maliciously., which is supposed to be picked up by algorithms or AI.

17. Section 3 of the amendment talks about "successive acts" of self harm, which cumulatively reach the threshold of GBH. "Successive" isn't defined but could mean three or more, which could be an arbitrary metric that platforms grab hold of to determine if something constitutes an offence.

18. Succession doesn't necessarily equate to intention as content could be shared for discussion and even to condemn the content, the acts of which would not be sufficient in a court of law but if it was to get to that stage , it would involve personal stress and expense.

19. The offence in section 185(7) may be committed by forwarding another person's direct message or sharing another person's post, a common error if someone shares that with a friend to say, "Look, this is what I just received. Isn't it awful."

20. The whole clause has a very broad potential application; consider section 185(4) which states that the person referred to in subsection 1 a) and b) need not be a specific person or a class of persons known to or identified by the defendant. The difficulty is with regards to proving intention when two concepts rub up against each other, that there could be intent to harm with no particular person in mind.

21. Section 185(5) says an offence can happen even if serious harm does not occur.

22. Section 185(6) extends liability so if someone encourages another to assist with self harm, then that first person is liable for the same act as the other.

Prior restraint and freedom of expression

23. The Bill tackles illegal content – content (words, images, speech or sounds) which amount to a relevant offence. The priority offences that online platforms must deal with are detailed in Schedules 5, 6 and 7.

24. Online platforms would do so by implementing the measures in clause 9(2a). It asks platforms to take or use proportionate measures relating to the design or operation of the service to *prevent individuals from encountering* priority illegal content by means of the service.

25. Two other relevant provisions of the Bill relate to mitigating and managing the risk of a service being used for the commission of a priority offence. That risk would be identified in a risk assessment conducted by the online platforms and service providers.

26. Article 19 from the UN Charter and Article 11 of the European Convention on Human Rights hold that everyone has the right to freedom of expression, including the right to hold opinions without interference, and to seek, receive and impart information and ideas through any media regardless of frontiers.

27. "Without interference" is the critical language for freedom of expression. It was written before automated systems came into the world and was generally understood to be interference by the state; in this instance interference is by a private actor at the request of the state.

28. Platforms already interfere and issue notices stating that content doesn't comply with their own terms and conditions/community standards before it is removed.

29. The Bill does allow complaints about the use of proactive technology - AI-driven content moderation technology implementing prior restraint – except users probably wouldn't know prior restraint has happened to affect that complaint.

30. A broad complaints procedure is included in the Bill but designated to Ofcom, as the regulator, to define.

31. The Bill should include a more detailed complaints procedure, including that users should have a right to be notified when the content is taken down, should be told what is being taken down and the reason for doing so and that users have the right to appeal to a judicial authority.

32. Since the roundtable , we have obtained a legal opinion from Dan Squires KC of Matrix Chambers, which confirms the points we addresses in this meeting. The opinion can be found via this link.

Observations

33. Nothing in the Bill suggests that take downs that might contain evidence of a crime will be accessible to relevant parties (forensic architects, prosecutors etc), which as seen during the Syria crisis when evidence of war crimes were wiped from social media.

34. We have to make assumptions about what the platform will deem, for example, terrorism offences as per schedule 5, and take into account the caution that will likely be exercised by social media platforms. Some foreseeability can be taken from the broader approach to terrorism labelling by the government for which the "Stansted 15" (protesting deportations) and environmental activists can be examples.

35. An overly cautious approach of social media companies would most likely result in takedowns of content opposing government policy; if there is plenty of content left online that is pro-government then there will be a one-sided dynamic in the public debate playing out online.

36. We can extrapolate from the presented examples and see implications for youth culture, particularly when their creative content – lyrical content and music that is seen as controversial (e.g. drill) but that is also lucrative and a form of expression – is already policed. A Metropolitan Police project seeks to take

down content it believes promotes violence including music, although recently, there is inclination to mine that content for evidence of gang behaviour used as digital evidence in criminal trials (a censured and racialised practice).

37. Others have noted it is more beneficial to leave content up as evidence if it shows acts of personal abuse etc.

38. Similarly, with regards to surveillance of public protests, the police prefer keeping content up to understand the networks people are part of, for example, to seek an injunction to material going on the internet.

39. While the Online Safety Bill refers to the older Public Order Act from 1986, the new Public Order Act 2023 has introduced serious disruption prevention orders – banning orders on protesting, using the internet or meeting with certain people – based largely on the content that people are posting online.

40. It's worth being aware that the "Henry VIII" clauses in the Bill will allow changes to come in with very little parliamentary scrutiny thus the resulting legislation could keep evolving though secondary legislation.

41. In addition to freedom of expression, you have the right to freedom of assembly, which is not something that only happens on the streets but also can apply online. This online aspect is being increasingly acknowledged in human rights standards with relevant guidance discussed by the Venice Commission. The UK has put its name on these standards and against the use of such restrictions online.

42. Automated decision-making will also involve coded bias. Using the example of Twitter, most takedown decisions, it is argued, have benefitted women or addressed hate speech against women. If you think about how AI will interpret pictures of a rifle club, consider if those in the images are people of colour and if that will mean those images are more likely to be considered terrorist content.

Conclusions

43. Transposing legislation of serious offences to the online sphere is very difficult.

44. There is no way platforms can tackle their obligations without using AI-driven content moderation systems. However, it lacks the ability to pick up the nuances and context involved in determining intent or other aspects of an offence.

45. Platforms will likely jump on arbitrary metrics, likely leading to unjust takedowns.

46. What will happen to any content that comprises "evidence"of crimes is a question still to be answered.

47. There is a concern important debate will be shut down and that what is left online is one-sided, devoid of any opposition or accountability for state policies and positions.

48. People's rights to freedom of expression will be infringed. Some people may experience a restriction on livelihoods.

49. Law enforcement access to evidence would also be disrupted

50. Taking down evidence could also increase harm if people are not receiving information for example, of death threats, content that amounts to stalking behaviour or other and as such they are not being warned that the threat exists.

51. The scope and scale of what the bill could cover could be endless because statutory instruments could be used by the government to add new provisions, potentially without any scrutiny. .

Emerging questions

52. Even if content contains evidence of an offence, doesn't the British public have a right to know?

53. Law in the counter-terrorism space is premised on broad definitions that cannot be effectively interpreted in the real world; so, how do we expect social media companies to effectively interpret the law?

54. If content is taken down, would police, prosecutors or journalists have access to that evidence?

55. What are the jurisdictional implications? For example, if content is taken down in one jurisdiction, is it taken down in all? Are there cross-border effects? E.g. If content that was lawful in South Africa, was uploaded from South Africa onto a UK platform, would it be illegal under the new Bill? Would there be cross-jurisdictional effects for content uploaded in South Africa, that was lawful in South Africa, but illegal under the Online Safety Bill in the UK?

56. In cases of abuse or stalking, for example, would the target of the content want the content left up or taken down?

57. What data will the AI-driven technology be trained on?

58. Will removal of “evidence” or content be welcomed by the police and are the provisions of the Bill well understood by the police forces?

Further Reading

Online Safety Bill as amended on Report Stage
https://bills.parliament.uk/publications/52368/documents/3841

Legal Advice on Prior Restraint Provisions in the Online Safety Bill

Facebook Oversight Board Decision: Geobbels https://oversightboard.com/decision/FB-2RDRCAVQ/#

Draft Online Safety Bill https://bills.parliament.uk/bills/3137