Index on Censorship

Response

Ofcom online safety call for evidence: First phase of online safety regulation

15 September 2022

Q1. Please provide a description introducing your organisation, service or interest in Online Safety.

Index on Censorship is a nonprofit organisation that campaigns for and defends free expression worldwide. We work tirelessly to protect freedom of expression and combat censorship on a global scale and publish work by censored writers and artists, promote debate, and monitor threats to free speech.

Index's aim is to raise awareness about threats to free expression and the value of free speech as the first step to tackling censorship. We have experience of dealing with users whose content has been restricted by content moderation systems and have published research which has informed our responses.

We are answering questions 11 and 18, and are happy to share any more information and reports with Ofcom as needed.

Section 5: Moderation

Q11. Could improvements be made to content moderation to deliver greater protection for users, without unduly restricting user activity? If so, what?

At present, the focus on content moderation in the Bill falls under a duty for platforms to censor content that is illegal or legal but harmful to 'an appreciable number of adults'. However, this vague concept will incentivise tech companies to use crude algorithms due to the sheer level of content they need to moderate. Companies will over-censor legal content to avoid fines of up to 10% of their global revenue. This measure will create a censorious UK internet, without actually making anyone safer from harm.

Algorithmic content moderation often disproportionately affects certain groups as it has no regard for context or nuance and will simply remove anything that it flags as 'harmful', even if it is not. These systems are still a long way away from being able to understand the subtleties in what content should be removed and are often developed without minority groups in mind. We have already seen this happen with Tumblr's system for identifying and removing adult content, introduced in December 2018, whichroutinely misclassified innocuous material, with content by LGBTQ+ users particularly penalised[1]. Studies from the University of

[1] Bright (2018), Matsakis (2018)

Washington have also already found that algorithms disproportionately identify posts written in African American Vernacular English as “rude” or “toxic,” reflecting and amplifying racial bias in AI.[2]

In 2016 a Swedish court case[3] was concluded against a former Syrian rebel who had taken part in the killing of seven captured Syrian soldiers. The court relied on content published on Facebook and Twitter to identify the time when and place where the soldiers were captured, as well as the fact that only 41 hours had passed between their capture and execution. Facebook was contacted by prosecutors in order to verify the content’s metadata.

As a charity that works to defend freedom of expression worldwide, we are well aware of the vital role that online content can play in highlighting and documenting crimes globally. Social media content has become a vital tool in identifying and prosecuting criminals, but over-censorship through algorithms risks removing crucial evidence of harm before users are able to report it to authorities. Social media content posted by perpetrators, victims, and witnesses to abuses, has become increasingly central to some prosecutions of war crimes and other international crimes, including at the International Criminal Court (ICC) and in national proceedings in Europe[4]. This also includes evidence of terrorism atrocities and rape: 23% of the Syrian War Crime Archive has already been deleted by platforms and will not be analysed and the perpetrators will not be prosecuted.[5]

To avoid a potential loss of vital evidence and deliver greater protection for users, we recommend that Ofcom mandates **the establishment of a Digital Evidence Locker.** Category 1 platforms should be required to archive and securely store all removed content from online publication alongside its reasoning for removal for a set period of time akin to HMRC requirements on public companies to keep financial records for 6 years from the end of the last company financial year they relate to. This would ensure that:

1) Ofcom, as the regulator, as well as parliament and third party researchers would be able to transparently audit content deletion and spot, over censorship, bias, or unintended consequences.
2) Crucial criminal evidence is not permanently lost and criminals are brought to justice by the police.
3) Users can adequately report crimes online and be aware of any threats made towards them.

---

[2] Bloch-Wehba (2020)

[3] NYT: Syrian Rebel Gets Life Sentence for Mass Killing Caught on Video (2018)

[4] Human Rights Watch report (September 2020)

[5] Reuters, ‘Lost memories’: War crimes evidence threatened by AI moderation (June 2020)

Further, to ensure that public interest materials are not inadvertently removed, Ofcom must **increase the thresholds for content moderation**[6] by matching the threshold for defining illegal content to criminal investigations and prosecutions offline. This will ensure that there is a level of consistency between speech that is criminalised through the courts and the standards applied by service providers online. The Bill should include a requirement for human moderation and particularly legal expertise on crime which will ensure that only content that is actually illegal is removed from platforms.

Q18. Are there any functionalities or design features which evidence suggests can effectively prevent harm, and could or should be deployed more widely by industry?

At present, **the Bill threatens to undermine end-to-end encryption** by proposing that accredited technology can be used to scan both publicly and privately shared content for anything related to terrorist activities. The scanning of everyone's private messages is a huge violation of privacy and completely disproportionate given existing policing and investigatory powers.

End-to-end encryption has its use for all of us: protecting us from hackers, safely sending family photos, and sharing personal information like medical history or bank details. **Doing this safely wouldn't be possible without encryption.**

1) Accessing encryption is also essential for journalists and whistleblowers not only for their work, but for their survival. For years, Index on Censorship has supported dissidents, journalists and activists by training them to use encryption and encrypted communication apps.

2) Encryption protects the collective security of UK citizens online from cyber attacks by criminals and states hostile to our democracies. But this Bill will push the UK away from Five Eyes cyber security standards and could open up UK citizens' private communications to foreign hostile states. The former head of GCHQ has already said that weakening security for everyone is not the solution. The UK Information Commissioner's Office (ICO) also intervened in the encryption debate with an unequivocal endorsement of end-to-end encryption.

[6] Gavin Millar KC A LEGAL ANALYSIS OF THE IMPACT OF THE ONLINE SAFETY BILL ON FREEDOM OF EXPRESSION (April 2022), pg 27