ICCL submission to the public consultation on regulation of online content

To the Department of Communications, Climate Action and Environment

Date 15 April 2019

Author Elizabeth Farries

The Irish Council for Civil Liberties (ICCL) would like to thank the Department of Communications, Climate Action & Environment for the opportunity to provide input towards the public consultation on the regulation of harmful content on online platforms.

In our submission, we briefly set out the issues and fundamental rights challenges facing state and corporate attempts to regulate content online and we also provide recommendations.

I. Fundamental rights implicated by harmful content regulation online

It is clear that our fundamental rights are implicated by state and corporate regulation of harmful content online, for example, our rights to privacy[1] and freedom of expression.[2] It is therefore important to observe at the outset that our rights are not changed or reduced online,[3] but rather apply to all forms of online communication.[4] Legislation in Ireland is required to conform with Ireland's human rights obligations under the Irish Constitution, the European Convention on Human Rights, and the international human rights treaties that Ireland has ratified.

Egregious circumstances including the exploitation of children, terrorism, or, more broadly, harmful content, are frequently cited reasons by states and corporations for limiting fundamental rights online. Mechanisms used and suggested to limit content online have included a combination of monitoring, reporting, pausing, reducing, removing, filtering, blocking, or censorship. Some of these terms signify the same actions, for example filtering and censoring, though the choice of language may soften rhetorically the implication for our rights that these actions might have.

Given the potential for rights limitations these content moderating actions entail, it is the position of the ICCL that states and corporations alike must comply with constitutional and international standards. While states may limit our enshrined rights only in exceptional

circumstances, they must still conform with the principles of legality, necessity, and proportionality.

II. Rights compliant online content moderation: identified difficulties

It is generally understood by rights groups in Ireland that online platforms have failed to regulate content in a manner that upholds fundamental rights in Ireland. Recently, various Private Members' motions have attempted to address this problem by introducing bills in the Dáil and the Seanad. Unfortunately however these bills, including the Digital Safety Commissioner Bill 2017, do not appear to have provided rights compliant methods or proposals for online content moderation.

For example, while the ICCL recognises the efforts of Deputy Ó Laoghaire to regulate harmful content by bringing forward the draft Digital Safety Commissioner Bill 2017, we have previously pointed out a number of difficulties with that bill. **In our Autumn 2018 joint submissions with CIVICUS to the Committee for Communications, Climate Action & Environment,[5] we described problems regarding the bill's incompatibility with international human rights standards on freedom of expression together with practical barriers to the bill's implementation.**

The ICCL otherwise observes several difficulties generally for rights compliant content regulation online:

Blanket monitoring is not rights compliant

Legislation or regulations permitting generalised monitoring of content based on the concern that it *might* be harmful could allow governments and corporate platforms to surveil people in Ireland in a manner that contravenes constitutional and human rights standards and the principles of legality, necessity and proportionality. Similarly, imprecisely drafted laws or regulations that do not intend to but nonetheless increase the chances of blanket surveillance would also run afoul of these standards.

Standardised definitions and removal procedures have been fallible

As legislative attempts in Ireland, at the EU level,[6] and also corporate filtering mechanisms have revealed, it can be very challenging to define terms for the purpose of regulating and removing content in a fair, consistent, or practical manner. As the explanatory note for this consultation points out, 'harmful content' as a term has been no exception. Further, in a now substantive body of evidence, we are also seeing that well resourced corporate platforms have been confronting the issue of ethical removal procedures for harmful content in a concerted and technically sophisticated way for several years. They continue to fail.[7] There are ever increasing examples of what has been in effect platform censorship rather than harmful content removal.[8]

The major barrier to rights compliant standardised term definitions or removal systems thus far have been problems in accuracy. Experts point out that, apart from settling on

agreed definitions, standardised content monitoring by either humans or algorithms are inevitably inaccurate - rights compliant material is often wrongfully removed and rights infringing material is often left up.[9]

Accuracy problems - the human element

On the human side, it is unrealistic to expect either corporate or government regulator employees to identify infringing content without a significant error rate. Mark Zuckerberg points out that 'The vast majority of mistakes [Facebook makes] are due to errors enforcing the nuances of our policies rather than disagreements about what those policies should actually be.'[10]Corporate platforms might be understandably accused of ulterior motives including profits attached to leaving content up.[11] However, it is also not clear how a team of human moderators at government regulatory levels might also receive adequate training to engage in rights balancing assessments historically reserved for the judiciary.

Accuracy problems - the machine element

On the machine side, algorithmic solutions to fundamental rights infringements have also been criticised by technical experts for having the unintended consequence of unduly limiting those rights. Automatic filters have not worked to date particularly because they can't detect context. Facebook admits this, noting that only 52% of hate speech is identified proactively.[12] Filters are also prohibitively expensive for any but the largest online platforms. In 2016, Google issued a report stating that YouTube's ContentID cost $60 million.[13] This number has likely continued to increase significantly and thus also raises for the ICCL a question regarding whether there is sufficient available government resourcing to assume these methods.

The legacy of state surveillance and censorship

There is also a legacy relating to the historical use of surveillance technologies deployed by states in a manner that disproportionately interferes with fundamental rights. These rights impacts cannot be disregarded particularly when they involve marginalized and discriminated communities. Surveillance tactics for example have been used by state policing institutions on social media platforms as tools of political and ideological persecution. The state goal has been to silence dissent, disrupt organised protest, crack down on social movements, and discredit protest leaders and social demands.[14] There are important questions raised by this legacy regarding how the Irish government might best respond to online content concerns without resorting to online surveillance.

Systems design and value-based solutions

Whether algorithmic design or predictive data can at some point effectively respond to the issue of content moderation is still being explored. There may be scope in the future for filters that are self-appointed and directed (as opposed to operated by an external authority including state regulators or corporate platform).[15] Self-appointed filtering would permit end users to decide what content we might see online. Such mechanisms might

include what one design expert has called ‘repository invitations’, a method whereby an internet user can't add another to a project without that user’s consent.[16]

These designs are nascent and still exploratory but redirect the conversation to the importance of users deciding for ourselves what we want our internet and online platforms to look like. A question that is absent from the consultation is one of values: will we support online spaces that are heavily monitored and tightly regulated by mediators who, to date, have often applied rights balancing analysis incorrectly? Or will we support online spaces that are free, secure, and self-actualising via the user's own discretion to control what content we have exposure to?

Transparency

One solution to the moderation problem proposed by the United Nations Special Rapporteur on Freedom of Expression includes ‘radical transparency’ for corporate platforms and states.[17] The Special Rapporteur’s brand of transparency has also been promoted by civil rights advocates.[18] It requires at a minimum full disclosure of the moderating entity of the rules used to moderate content and how those rules are applied, together with appeals processes and accountability for wrongful takedown.

Transparency to this level of disclosure would assist state, treaty body mechanisms and human rights advocates in better understanding the strengths and weaknesses of content moderation programs towards designing more effective and rights compliant programs. See for example the repeated requests by Amnesty International for Twitter to publish data on the abuse perpetrated on their platform, and see also the company’s failure to do so thus far. Amnesty International correctly observes that this opacity hides the extent of the content moderation problem and makes it difficult to design effective solutions.[19]

III. Recommendations

In this challenging and still evolving landscape of online content moderation, the ICCL stresses that it is important to centre rights based analysis in proposed solutions. We request that decision makers not engage in what some have called ‘magical thinking’ by continuing to rely on content moderation mechanisms that have been proven to be both ineffective and harmful. With this preface we provide the following four recommendations:

Rights compliant moderation

As the Special Rapporteur on Freedom of Expression has made abundantly clear: ‘States should only seek to restrict content pursuant to an order by an independent and impartial judicial authority, and in accordance with due process and standards of legality, necessity and legitimacy.’[20] The ICCL supports this position while noting that the precise mechanisms for achieving this have not yet been identified. Given this lack of clarity, it is the position of the ICCL that content moderation should adhere absolutely to a rights-based approach which complies with constitutional and human rights standards. Legal ambiguities relating to moderating online content should be resolved in favour of respect for freedom of expression, privacy, and data protection principles.

Transparency

The ICCL supports the recommendations of the Special Rapporteur[21] and also the Santa Fe principles[22] for explicit transparency. We assert transparency is essential for both corporate platform and state content moderation. Transparency includes at minimum full disclosure of the rules used to moderate content and how those rules are applied together with functional appeals processes and accountability for wrongful takedown.

Harmful content definitions

States must clarify definitions of harmful content so that they may be subject to a rights balancing analysis. It is unlikely that states can define harmful content to a level of specificity that avoids the need for an independent and impartial judicial authority to evaluate individual circumstances when applying this definition.

Blanket monitoring

Blanket monitoring, particularly by cloud services and infrastructure, software and platform services, should be prohibited in order to protect fundamental rights. This includes prohibiting automated monitoring tools including filters that are used to surveil content generally and indiscriminately online.

These submissions were drafted by Elizabeth Farries, Surveillance and Human Rights Program Manager with the ICCL (https://www.iccl.ie) and INCLO (https://www.inclo.net). For further inquiries please contact info@iccl.ie.

About the ICCL

The Irish Council for Civil Liberties is Ireland's leading independent human rights organisation. We monitor, educate and campaign in order to secure full enjoyment of rights for everyone. Founded in 1976 by Mary Robinson and others, the ICCL has played a leading role in some of the most successful human rights campaigns in Ireland. We have previously given submissions to Oireachtas Committees on digital rights and privacy questions, including the *Communications (Retention of Data) Bill* and *The Public Services Card.* We have also raised human rights questions guiding legislative responses to recent 2018 social media scandals including data protection breaches and degrading content.

Please see our previous submissions on this topic at https://www.iccl.ie/wp-content/uploads/2018/11/Digital-Safety-Commissioner-Bill-2017-ICCL-CIVICUS-Submissions.pdf

1 Our right to privacy is protected by Article of 12 of the Universal Declaration of Human Rights (UDHR), Article 17 of the International Covenant on Civil and Political Rights (ICCPR), Article 8 of the European Convention on Human Rights (ECHR), and Article 7 of the Charter of Fundamental Rights of the EU (EU Charter). In correlation, our personal data is

also protected under Article 8 of the EU Charter. In the Irish Constitution, a right to privacy has also been identified as one of the unenumerated rights stemming from the wording of Article 40.3; see *Cullen v. Toibín* [1984] ILRM 577.

2 Similarly, our right to freedom of expression is protected by Article 19 of the UDHR, Article 19 of the ICCPR, Article 10 of the ECHR, Article 11 of the EU Charter, and Article 40(6)(1)(i) of the Irish Constitution.

3 The United Nations Human Rights Council has stated that the same rights people have offline must also be protected online. This is particularly true for freedom of expression, which is applicable regardless of frontiers and through any media of one's choice, in accordance with articles 19 of the UDHR and the ICCPR. See UN Doc A/HRC/32/L.20.t, available at: https://documents-dds-ny.un.org/doc/UNDOC/LTD/G16/131/89/PDF/G1613189.pdf?OpenElement

4 The right to freedom of expression for example applies to all forms of electronic and Internet-based modes of expression. See UN Human Rights Committee, General Comment No.34 on Article 19: Freedoms of opinion and expression, CCPR/C/GC/34, (2011), available at: https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf

5 Please see the ICCL's previous joint submissions with CIVICUS, available at: https://www.iccl.ie/wp-content/uploads/2018/11/Digital-Safety-Commissioner-Bill-2017-ICCL-CIVICUS-Submissions.pdf

6 See the parallel and challenged attempts of the European Parliament Committee on Civil Liberties, Justice and Home Affairs (LIBE) to define and moderate terrorist content online. EDRI, 'Terrorist Content Regulation: Successful 'damage control' by LIBE Committee' (*EDRI*, 08 April 2019), available at: https://edri.org/terrorist-content-libe-vote/

7 See many strong examples of this failure outlined by Corynne McSherry and Gennie Gebhart, 'Mark Zuckerberg Does Not Speak for the Internet' (*Electronic Frontiers Foundation*, 1 April 2019), available at: https://www.eff.org/deeplinks/2019/04/mark-zuckerberg-does-not-speak-internet

8 The Electronic Frontiers Foundation categorises these examples and also provides anecdotes: 'We've seen prohibitions on hate speech used to shut down conversations among women of color about the harassment they receive online; rules against harassment employed to shut down the account of a prominent Egyptian anti-torture activist; and a ban on nudity used to censor women who share childbirth images in private groups. Museums have had works of art taken down for "suggestive content." And we've seen false copyright and trademark allegations used to take down all kinds of lawful content, including time-sensitive political speech.' See Corynne McSherry and Gennie Gebhart, 'Mark Zuckerberg Does Not Speak for the Internet' (*Electronic Frontiers Foundation*, 1 April 2019), available at https://www.eff.org/deeplinks/2019/04/mark-zuckerberg-does-not-speak-internet; see also the University of Toronto's Citizen Lab results of their investigation into how internet internet-filtering technology in Canada is being used in various countries to censor access to news, religious content, LGBTQ+ resources, and political campaigns. Jakub Dalekhttp et al, 'Planet Netsweeper, Executive Summary' (*Citizen Lab*, 25 April 2018), available at: https://citizenlab.ca/2018/04/planet-netsweeper/

9 See Daphne Keller, 'Problems with filters in the European Commission's platforms proposal' (*Center for Internet and Society at Stanford Law Schoo*l, 05 October 2017), available at: http://cyberlaw.stanford.edu/blog/2017/10/problems-filters-european-commissions-platforms-proposal. Keller notes that 'errors include both false positives—removing lawful content—and false negatives—leaving infringing content up.'

10 Mark Zuckerberg, 'A Blueprint for Content Governance and Enforcement' (*Facebook*, 15 November 2018), available at: https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/

11 BBC News, 'Facebook moderators "keep child abuse online"' (*BBC News*, 17 July 2018) available at: https://www.bbc.com/news/technology-44859407

12 Mark Zuckerberg, 'A Blueprint for Content Governance and Enforcement' (*Facebook*, 15 November 2018), available at: https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/

13 Google (2016). *How Google Fights Piracy*, available at: https://drive.google.com/file/d/0BwxyRPFduTN2cl91LXJ0YjlYSjA/view

14 The ICCL, as a member of the International Network of Civil Liberties Organisations (INCLO), has also written about this surveillance problem in our March 2019 submissions to the UN Special Rapporteur on the rights to freedom of peaceful assembly and of association to inform his thematic report - The rights to freedom of peaceful assembly and of association in the digital age. INCLO will be publishing our report in June 2019.

15 Mike Masnick, 'Platforms, Speech And Truth: Policy, Policing And Impossible Choices' (*techdirt*, 9 August 2019), available at: https://www.techdirt.com/articles/20180808/17090940397/platforms-speech-truth-policy-policing-impossible-choices.shtml

16 Bits of Freedom, 'ENDitorial: Can design save us from content moderation?' (*ENDitoria*l, 16 May 2018), available at: https://edri.org/enditorial-can-design-save-us-from-content-moderation/

17 UN Doc A/HRC/38/35 (18 June−6 July 2018).

18 See in particular 'The Santa Clara Principles On Transparency and Accountability in Content Moderation', available at: https://santaclaraprinciples.org. The principles state that, at minimum, companies should (1) publish the numbers of posts removed and accounts permanently or temporarily suspended due to violations of their content guidelines; (2) provide notice to each user whose content is taken down or account is suspended about the reason for the removal or suspension; and (3) provide a meaningful opportunity for timely appeal of any content removal or account suspension. It is the position of the ICCL that states should be held to an equally high transparency standard.

19 Amnesty International (2018), '#TOXICTWITTER. Violence and abuse against women online', available at: https://www.amnesty.ca/sites/amnesty/files/%23TOXICTWITTER%20report%20EMBARGOED.pdf

20 Freedex, 'The Special Rapporteur's 2018 report to the United Nations Human Rights Council is now online' (*A Human Rights Approach to Platform Content Regulatio*n, 6 April 2019), available at: https://freedex.org/a-human-rights-approach-to-platform-content-regulation/

21 UN Doc A/HRC/38/35 (18 June−6 July 2018).

22 'The Santa Clara Principles On Transparency and Accountability in Content Moderation', available at: https://santaclaraprinciples.org. The principles state that, at minimum, companies should (1) publish the numbers of posts removed and accounts permanently or temporarily suspended due to violations of their content guidelines; (2) provide notice to each user whose content is taken down or account is suspended about the reason for the removal or suspension; and (3) provide a meaningful opportunity for timely appeal of any content removal or account suspension. It is the position of the ICCL that states should be held to an equally high transparency standard.