Committee of experts on
freedom of expression and digital technologies
(MSI-DIG)

COUNCIL OF EUROPE

CONSEIL DE L'EUROPE

18 March 2021 MSI-DIG(2021)04

Draft

Guidance Note on best practices towards effective legal and procedural frameworks for self-regulatory and co-regulatory mechanisms of content moderation

Introduction

1. Content moderation (here understood in a broad sense, encompassing also content curation) is increasingly used as a means to address a variety of issues arising as a result of users' activity in the online environment. It poses particular challenges because, due to evolving technologies, the possibilities they offer and the constant evolution of human behaviour in the online environment, "[t]here is no end-state of content moderation with stable rules or regulatory forms; it will always be a matter of contestation, iteration and technological evolution."[1]

2. Self-regulation and co-regulation, often referred to as two discrete approaches to content moderation, in fact represent variable benchmarks on the same continuum, with a purely self-regulatory approach on one end, and a purely regulatory approach on the other, depending on the degree of state implication in the process.[2]

3. Co-regulation involves a greater degree of engagement of the state and is often related to the achievement of pubic policy objectives, while self-regulation would usually be introduced by internet intermediaries independently, for reasons directly related to their business model (hereafter – business purposes).

4. However, at any level content moderation inevitably involves human rights considerations. States should be mindful that their positive and negative human rights obligations, including with regard to the rights to freedom of expression, privacy, freedom of assembly and association, equality and non-discrimination, and the right to an effective remedy, arise also from content moderation performed within self- and co-regulatory frameworks.[3]

5. The purpose of this Guidance note is to provide practical guidance to member States of the Council of Europe for policy development, regulation and use of content moderation in the online environment in line with their human rights obligations under the European Convention on Human Rights. The Guidance note is also addressed to internet intermediaries who have human rights responsibilities of their own).[4]

6. To this end, the Guidance note seeks to discern best practices towards effective legal and procedural frameworks for self-regulatory and co-regulatory mechanisms of content moderation.

7. The Guidance note builds on an impressive body of work on various aspects of content moderation already produced by a range of institutions, as referenced in the annexed Explanatory

---

[1] Evelyn Douek, "The limits of international law in content moderation," UCI Journal of International, Transnational, and Comparative Law", December 2020, p. 9. https://ssrn.com/abstract=3709566 (last accessed 21 January 2021).

[2] C.T. Marsden, "Internet co-regulation and constitutionalism: Towards European judicial review International Review of Law Computers & Technology" 26(2):211-228, November 2012, https://www.researchgate.net/publication/254294662_Internet_co-regulation_and_constitutionalism_Towards_European_judicial_review (last accessed 10 June 2020).

[3] *Peck v. the United Kingdom*, no. 44647/98, 2003-I, paras. 108 and 109.

[4] See UN "Guiding Principles on Business and Human Rights: Implementing the UN 'Protect, Respect, and Remedy' Framework" and Recommendation CM/Rec(2016)3 of the Committee of Ministers to member States on human rights and business.

memorandum. Of particular note is the work of the Council of Europe,[5] [6] the United Nations, [7] [8] [9] academic institutes,[10] [11] [12] and a variety of NGOs.[13] [14] [15] [16] [17] [18] The Explanatory memorandum is meant to support and explain the provisions, concepts and terminology contained in the Guidance note and to be used extensively for reference in the process of its implementation.

General considerations for good policymaking

8. States should be mindful that content moderation raises complex and unresolved issues surrounding jurisdiction. The transborder nature of the online environment should be duly taken into account in any regulatory decisions related to content moderation, as well as in impact assessments.[19] [20]

9. Given the fast and constant evolution of the online environment, content moderation policies require regular review. States should ensure appropriate supervision and timely adaptation of relevant policies and regulatory frameworks.

10. Content moderation is used to address a wide range of public policy problems, from various forms

[5] Recommendation CM/Rec (2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries.

[6] Recommendation CM/Rec (2020)1 of the Committee of Ministers to member States on the human rights implications of algorithmic systems.

[7] United Nations Human Rights Council, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue, 2011. https://www2.ohchr.org/english/bodies/hrcouncil/docs/17session/A.HRC.17.27_en.pdf (last accessed 23 September 2020).

[8] David Kaye, "A New Constitution for Content Moderation," medium.com, June 2019, https://onezero.medium.com/a-new-constitution-for-content-moderation-6249af611bdf (last accessed 23 September 2020); Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression of 6 April 2018, A/HRC/38/35, http://daccess-ods.un.org/access.nsf/Get?Open&DS=A/HRC/38/35&Lang=E (last accessed 23 April 2021).

[9] Broadband Commission for Sustainable Development, Report "Balancing Act: Countering Digital Disinformation while respecting Freedom of Expression", 2020, https://en.unesco.org/publications/balanceact (last accessed 23 April 2021).

[10] https://www.ivir.nl/publications/technology-and-law/

[11] Luca Belli et al, Platform Regulations: "How platforms are regulated and how they regulate us," December 2017, https://direitorio.fgv.br/publicacoes/platform-regulations-how-platforms-are-regulated-and-how-they-regulate-us (last accessed 19 February, 2021).

[12] https://cyberlaw.stanford.edu/focus-areas/intermediary-liability.

13 AccessNow, "26 recommendations on content governance," 2020. https://www.accessnow.org/cms/assets/uploads/2020/03/Recommendations-On-Content-Governance-digital.pdf (last accessed 23 September 2020).

[14] Article 19, "Side-stepping Rights: Regulating Speech by Contract," 2018, https://www.article19.org/wp-content/uploads/2018/06/Regulating-speech-by-contract-WEB-v2.pdf (last accessed 4 January 2020)

[15] See santaclaraprinciples.org (last accessed 23 September 2020).

[16] See platformregulation.eu (last accessed 23 September 2020).

[17] See https://edri.org/?s=content+moderation(last accessed 23 September 2020).

[18] Meedan, "Content Moderation Toolkit, November 2018, https://meedan.com/reports/content-moderation-toolkit/ (last accessed 19 January 2020).

[19] For example, content that is prohibited in one jurisdiction can be removed globally, due to orders implemented in that jurisdiction or, to an extent, restricted only in that jurisdiction (e.g., Google globally removes content subject to procedurally correct complaints under the US Digital Millennium Copyright Act, but limits the restrictions imposed under the so-called "right to be forgotten" ruling of the Court of Justice of the EU (Case C-131/12) to searches carried out under its EU operations).

[20] Similarly, liability rules imposed in larger jurisdictions, or jurisdictions where intermediaries are based, can have an impact on freedom of expression and the right to receive and impart information on a global level. See, for instance, Dan Jerker B. Svantesson, "Internet and Jurisdiction Global Situation Report 2019," https://www.internetjurisdiction.net/uploads/pdfs/GSR2019/Internet-Jurisdiction-Global-Status-Report-2019_web.pdf (last accessed 28 September 2020), and Internet & Jurisdiction Project, "I&J Outcomes: Mappings of Key Elements of Content Moderation," June 2020, https://www.internetjurisdiction.net/news/i-j-outcomes-mappings-of-key-elements-of-content-moderation (last accessed 28 September 2020).

of criminal behaviour to content that is not illegal, as well as content moderated by internet intermediaries for business purposes (such as addressing "spam"). It is therefore fundamental to good policymaking in this area that policies are designed with a clear understanding and recognition of the nature of the content being addressed and the accountability of internet intermediaries in the decision-making process.

11. Similarly, it is important that policymaking take account of the different challenges of self- and co-regulation in different contexts. These challenges vary depending, for example, on the *nature and* scale of hosted content and its reach, and are not the same for internet intermediaries and for media outlets.

12. States have a variety of positive and negative obligations in this context. They must create sufficiently developed regulatory frameworks for content moderation that upholds the exercise and enjoyment of human rights of internet users, including victims of illegal content. States must protect the rights to freedom of expression, privacy, freedom of assembly and association, equality and non-discrimination, the right to an effective remedy and other human rights of everyone within their jurisdiction when these rights are affected by content moderation.

13. The principle of proportionality, well developed in the case law of the European Court of Human Rights, requires that any restrictions to human rights be the least restrictive necessary.[21] Beyond the binary choice between deletion or not, content moderation offers a range of tools (such as temporary or permanent demotion, demonetisation or tagging as problematic) that should be considered by policy-makers and put into use with due regard to the nature of content addressed.

14. In relation to content moderation implemented to address public policy objectives, this requires states to:

    a. be clear about the nature of the problem(s) being addressed;

    b. ensure the predictability of the measures implemented;

    c. ensure that legal and regulatory frameworks do not result in overcompliance or discriminatory implementation;

    d. identify the rights at risk in each context and be clear about how they will be upheld;

    e. be clear about the nature of the regulatory approach to content moderation by indicating the degree of state engagement;

    f. actively learn from experience of similar self- and co-regulatory schemes, in order to maximise effectiveness and minimise unintended consequences, and

    g. ensure that all necessary transparency data is generated and made public to ensure accountability for, and identification and rectification of, problems.

Defining the problem

15. Content that may be subject to content moderation for public policy reasons varies considerably. Each category of content represents a separate public policy challenge, with specific characteristics. Respective state policies should aim to carefully distinguish between different

[21] See, for instance, *Autronic AG v. Switzerland*, no. 12726/87, 22 May 1990, para. 61; *Weber v. Switzerland*, no. 11034/84, 22 May 1990, para. 47; *Barthold v. Germany*, no. 8734/79, 25 March 1985, para. 58; *Klass and others v. Germany*, no. 5029/71, 6 September 1978, para. 42; *Sunday Times v. the United Kingdom*, no. 6538/74, 26 April 1979, para. 65; *Observer and Guardian v. the United Kingdom, no. 13585/88*, 26 November 1991, para. 71.

categories of illegal content and develop targeted, efficient and proportionate responses, fitted to the characteristics of the concrete problem being addressed.[22]

16. To minimise the risk of hasty, ineffective, counterproductive or disproportionate responses to newly arising challenges, States should aim to develop a clear, public methodology for categorising different types of content and developing adequate and human rights compliant responses. Such a public methodology should be made available by both states and online platforms in a transparent and easily accessible manner.

17. Continuous assessment of the evolution of the problem based on a specific category of illegal content, its drivers and impact on society is essential to ensure proportionality, predictability and effectiveness of the means employed to address it.

18. In particular, it is crucial that any self- or co-regulatory obligation to fight serious crime (e.g., that constitutes a threat to human life, or child abuse) include obligations for states to take all necessary measures to address the offline component of the crime. It should never be possible to adopt a self- or co-regulatory approach in relation to such content without explicit reference to the expected engagement with law enforcement and other relevant state authorities.[23] Such measures should aim for maximum transparency and need to be calibrated very carefully in order to avoid unintended consequences for privacy and other human rights, as well as prevention of abuse of power by state authorities.

19. Where content moderation is used by internet intermediaries for business purposes, states should ensure that any restrictions resulting from it are clear, predictable and imposed in a non-discriminatory way, that does not lead to undue interference with human rights, and ensure adequate and accessible redress.

Ensuring predictability

20. Restrictions on human rights need to be predictable, in order to allow individuals to regulate their behaviour. This covers legal prohibitions of content and liability rules imposed on internet intermediaries by states. This also applies to the design, updating and implementation by internet intermediaries of their terms of service.

21. State policies on content moderation must be non-discriminatory and take into account the substantial differences in size and scale of internet intermediaries. States should avoid promoting approaches which impose on internet intermediaries disproportionate obligations, or that delegate to them decision-making to the detriment of democratically legitimated approaches, including by denying any stakeholder group the right to provide meaningful input.

22. Similarly, with regard to content moderation undertaken by internet intermediaries for business purposes, all necessary transparency and procedural safeguards need to be put in place to avoid and, if necessary, identify and remedy any discriminatory bias, particularly against vulnerable groups.

23. It is incumbent on the state to ensure a legal framework that is predictable for all concerned. If internet intermediaries are to be held liable for failing to remove illegal content, the rules

[22] For example, addressing content which is evidence of an offline crime (such as child abuse) requires a different response from content that is illegal due to its context (such as non-consensual publication of intimate images, known as "revenge pornography").

[23] For example, the German "Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken" / "Network Enforcement Law" ("NetzDG") requires internet intermediaries to store some associated data, when they remove content, in case this is needed for subsequent investigations.

concerning "knowledge" triggering that liability must be clear and proportionate, as must the rules prohibiting the content in question.

24. Predictability also requires the nature and extent of state involvement in the mechanism imposing the restriction on content to be clear. States should ensure that in all cases legal obligations and responsibilities, as well as operational roles and accountability requirements are clearly defined.

25. States should be mindful that the more urgent it is to engage in restrictions, the more important it is to ensure their effectiveness and accountability for them. Transparency, review and adjustment mechanisms are essential and should not be taken as secondary considerations because of urgency.

26. When internet intermediaries use specialised organisations (variously referred to as "trusted flaggers" or "priority flaggers")[24] as a filter through which to get more reliable reports of infringing content, specific transparency rules are needed for such initiatives, in order to ensure that no perverse incentives or conflicts of interest are accidentally created and that their level of effectiveness and trustworthiness remains consistently high. Use of trusted flaggers should not be mandatory and notices from trusted flaggers should not be considered "actual knowledge" of illegality of content. The status of "trusted flaggers" should be periodically, independently evaluated.

27. While "trusted flaggers" exist to assist with identification of content to be removed, consideration should be given to the fact that (apart from a number of informal arrangements) there is no mechanism for trusted groups to formally request the remedying of content moderation errors. The viability of "trusted de-flagger" systems should therefore be investigated.

Overcompliance and discrimination

28. States must ensure an appropriate balance of incentives for internet intermediaries and avoid regulation incentivising them to impose disproportionate restrictions. This can happen, for example, as a result of intermediary liability rules that are either too stringent or too vague. This is particularly relevant for content that is legal but possibly undesirable in a democratic society, where it is recognised that human rights must also be upheld. [25]

29. Decisions taken by human beings and by technological systems that they create are not infallible and have deliberate and accidental biases. State policies should require that enough data is made publicly available by internet intermediaries to ensure adequate, independent auditing capable of identifying any discriminatory or problematic approaches in content restriction decisions.

30. Furthermore, content moderation systems and associated complaints mechanisms have the potential to be abused ("gamed") by bad actors who can target speech from groups they oppose and have this speech restricted. This can happen as a result of malicious behaviour by individuals and also through coordinated behaviour (numerous individuals submitting individual complaints) and must be anticipated and mitigated. States should be mindful that, while dissuasive sanctions are essential to address such abuses, they cannot be relied on as a complete solution.

[24] EuroISPA, "Priority Flagging Partnerships in Practice", January, 2019, https://www.euroispa.org/wp-content/uploads/Hutty_Schubert_Sanna_Deadman-Priority-Flagging-Partnerships-in-Practice-EuroISPA-2019.pdf (last accessed 28 May 2020).

[25] In line with the case law of the European Court of Human Rights, ideas "that offend, shock or disturb the State or any sector of the population," must have a means to be expressed - see, among others, *Handyside v. the United Kingdom*, no. 5493/72, 7 December 1976, para. 49.

Affected rights

31. All content moderation policies should be assessed at the outset and on an ongoing basis with regard to the possible impact on the human rights they may restrict. Particular care should be exercised to ensure that human rights are protected, with restrictions being implemented in full respect of the European Convention on Human Rights.

32. Building on existing Council of Europe standards,[26] the right to effective remedy requires that individuals be informed precisely of the basis on which their content was removed or why their complaint did not lead to content being removed. It requires the right to accessible adjudication. While access to judicial authorities should always be possible, if requested by either party, states should support the making available of alternative dispute resolution mechanisms, innovative multistakeholder-designed arbitration solutions or e-courts, as appropriate.

33. Available redress should take into account that the cost to the injured party, and to society in general, may not be financial and, in the online environment, the mere publication of a correction may not adequately remedy the initial damage or infringement.

34. Due attention must also be given to the labour rights and mental health of all workers involved in manual review of content which may be shocking, disturbing or otherwise likely to have a psychological impact on the individuals concerned. This is particularly the case when internet intermediaries outsource this task to third parties, possibly based in other countries with different and less protective labour laws.

Nature of the regulatory approach

35. Decision-making by internet intermediaries in relation to content moderation can happen on a wide continuum, from decisions taken fully independently to decisions taken as a direct result of pressure by non-legal means from states, from media or civil society campaigns or as an intended or unintended consequence of intermediary liability rules for a wide range of offences. States should be mindful of and recognise that the extent of their implication in decision-making by internet intermediaries influences the scope of their related human rights obligations.

36. Because co-regulatory approach to content moderation implies a higher level of obligations on the state and allows the state to ensure a greater degree of inclusiveness, accountability and transparency, it is generally more suitable in contexts where matters of public policy are at stake. When engaging in co-regulation, States should use clear language and definitions regarding the nature of cooperation with internet intermediaries, its goals and targets, the responsibilities and obligations of all parties involved.

Characteristics of successful and failed approaches

37. Research on the characteristics of successful self-regulation in other fields, which can also be applied to co-regulation, have identified several key traits:[27]

    a. Transparency, particularly with regard to targets, balance of power and independence;

    b. Having clear, independently verified, objective benchmarks and targets;

[26] See section 2.5 of Recommendation CM/Rec (2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries.

[27] Sharma, L. L., Teret, S. P., & Brownell, K. D. (2010). "The food industry and self-regulation: standards to promote success and to avoid public health failures." American journal of public health, 100(2), 240–246. https://doi.org/10.2105/AJPH.2009.160960 (last accessed 8 October 2020)

c. Mandatory public reporting and independent appraisal and audits of adherence to codes and on progress towards goals being achieved, and the imposition of appropriate sanctions where non-compliance is identified;[28]

d. Ongoing independent oversight.

38. The experience showing which self- or co-regulatory approaches have been fully or partially successful, and which have not, should be regularly reviewed, in order to continually improve the effectiveness and human rights compatibility of self- and co-regulatory approaches to content moderation.

Transparency

39. The key benefit of self-regulation and co-regulation is their flexibility, which is particularly valuable in the continuously changing online environment, where newly arising issues require adequate and timely responses. States and internet intermediaries should be mindful of and acknowledge that without meaningful transparency, society loses this benefit of self- and co-regulation, while still incurring a reduction in accountability and democratic legitimacy that come as their cost.[29]

40. Transparency is essential for content moderation to respect human rights and to achieve its goals. It is needed:

- to be clear about the nature of the content moderation (whether it is implemented within a self- or co-regulatory approach and why, whether it is implemented directly or due to embedded incentives);
- to be clear about the problem being addressed and its targets;
- to be clear about the rights that are potentially restricted;
- to be clear about any fully or partly automated processes used in making content moderation decisions;
- to be clear, as necessary, what types of non-illegal content or behaviours are not permitted on the services of an internet intermediary;
- to ensure that the least restrictive alternative is being used when rights are being restricted;
- to identify and eliminate mistakes that lead to legitimate content being removed or illegitimate content being left online.

41. In order to ensure adequate transparency and auditing, it is crucial for internet intermediaries, in full respect of data protection law and principles, to preserve previously restricted content and the reasons why it was restricted.

42. States should be mindful that the speed and quantity of deleted content items do not necessarily indicate the efficiency of measures. Metrics should be oriented towards efficiency- that is indicative of the progress made in achieving concrete public policy objectives.

43. Similarly, an important value of transparency is being able to track problems across different internet intermediaries and over time. If the methodologies and reporting formats of internet

[28] European Commission, study on the "Effectiveness of self- and co-regulation in the context of implementing the Audiovisual Media Services Directive (AVMSD)", 2016, https://ec.europa.eu/digital-single-market/en/news/audiovisual-and-media-services-directive-self-and-co-regulation-study

[29] Without meaningful transparency, it is not possible to identify and assess changes nor make corresponding adjustments in policies. See also section 2.2 of Recommendation CM/Rec (2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries.

intermediaries do not permit such tracking, then this resource will not be available to policymakers.

44. For full and effective transparency, data should be provided with maximum levels of granularity and the most consistent methodology possible, across similar types of internet intermediaries and over time, allowing an effective analysis and evaluation of the content moderation methods applied, and should be made publicly available in clear language.

Draft

Explanatory memorandum to the Guidance Note on best practices towards effective legal and procedural frameworks for self-regulatory and co-regulatory mechanisms of content moderation

Contents

Key concepts

The purpose of this section is to introduce key concepts which will be used throughout the Explanatory Memorandum and in the Guidance Note.[30]

Censorship: Restriction of the use of certain images, words, opinions, or ideologies. The word is used here in the value-neutral English legal sense.[31]

Content curation: The process of deciding which content should be presented to users (in terms of frequency, order, priority, and so on), based on the business model and design of the platform.

Content moderation: The process whereby a company hosting online content assesses the [il]legality or compatibility with terms of service of third-party content, in order to decide whether certain content posted, or attempted to be posted, online should be demoted (i.e., left online but rendered less accessible), tagged as being potentially inappropriate or incorrect, demonetised,[32] not sanctioned or removed, for some or all audiences, by the service on which it was posted.

Co-regulation: Measures taken proactively by companies or sectors, in cooperation with or under supervision of states, to either:

- demonstrate compliance with a legal obligation or with non-binding agreements or codes or,
- regulate their activities, as a result of negotiation or cooperation with states, or as the result of encouragement from states. In the EU, it can be a "mechanism whereby a Community legislative act entrusts the attainment of the objectives defined by the legislative authority to parties which are recognised in the field".[33]

A co-regulatory model that provides a legal underpinning for a self-regulatory and demonstrably independent body can offer an approach which upholds international standards for freedom of expression.

While the EU Audiovisual Media Services (AVMS) Directive fails to define co-regulation, it does provide a description of what it is understood to be.[34]

---

[30] An extensive discussion on these and related concepts is, at time of writing, underway in the Internet Governance Forum Coalition on Platform Responsibility. See https://www.intgovforum.org/multilingual/content/glossary-on-platform-law-and-policy-terms (last accessed 15 January 2020) for more information.

[31] While this word is often used in a pejorative sense, it is used here in a neutral legal sense, as is common in English-speaking countries. For example, the remit of the Irish Board of Film Classification is established by the Censorship of Films Act 1923 and the Director of Film Classification is formally the "Official Censor." See http://www.ifco.ie/en/ifco/pages/legislation (last accessed 06 January 2020). The equivalent UK body was called the "British Board of Film Censors" until 1984.

[32] Some platforms remunerate users for uploaded content (allowing the user to "monetise" their content) and may remove such payments, in certain circumstances. See https://support.google.com/youtube/answer/6162278 (last accessed 13 October, 2020) for information about the Google/YouTube policy.

[33] 2003 Interinstitutional Agreement on Better Lawmaking, 2003/C 321/01, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.C_.2003.321.01.0001.01.ENG (last accessed 10 June 2020). (paragraph 18). Crucially in this context, this document also clearly stated that co- and self-regulation "mechanisms will not be applicable where fundamental rights or important political options are at stake" (paragraph 17). This instrument is no longer in force. Its successor does not mention co- and self-regulation.

[34] "Co-regulation provides, in its minimal form, a legal link between self-regulation and the national legislator in accordance with the legal traditions of the Member States. In co-regulation, the regulatory role is shared between stakeholders and the government or the national regulatory authorities or bodies.: Directive 2018/1808 of the on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive), recital 14. https://eur-lex.europa.eu/eli/dir/2018/1808/oj (last accessed 19 February, 2021).

Regulation: A state-imposed rule/legal obligation or the act of controlling something.

Self-regulation: Voluntary measures undertaken by companies or sectors without government encouragement. A recital of the EU AVMS Directive describes self-regulation as follows: "Self-regulation constitutes a type of voluntary initiative which enables economic operators, social partners, non-governmental organisations and associations to adopt common guidelines amongst themselves **and for themselves**. They are responsible for developing, monitoring and enforcing compliance with those guidelines."[35] (emphasis added)

Furthermore, in practice, the terms "self-regulation" and "co-regulation" have been used with a considerable degree of overlap and can be subdivided into several sub-categories.[36]

Content moderation raises particular challenges regarding the role and accountability of states.[37] These concerns arise from the fact that it is a task generally implemented by private parties, often to achieve public policy objectives. As a result, these explanations of concepts emphasise the fact that States are the main human rights duty bearers.

I Introduction

The internet has given us fantastic new opportunities to speak, to be heard and to organise. Indeed, it has created a wealth of new opportunities to exercise our human rights, including our rights to freedom of expression, freedom of assembly, freedom of thought and religion and others. However, unsurprisingly, it also creates opportunities for illegal or potentially damaging content or behaviour to be spread. Online services (such as social media platforms, where people post messages, articles, pictures, and so on), have a process of deleting, demoting or otherwise discouraging the spread of illegal or unwelcome content that is referred to as "content moderation".

A great deal of excellent research and analysis has already been done on the operational aspects of content moderation, by international organisations like the United Nations, non-governmental organisations individually and collectively, and academics. It is not the aim of the Guidance Note to duplicate or even to thoroughly catalogue this impressive body of work. Of note are:

- former UN Special Rapporteurs Frank LaRue[38] and David Kaye,[39]

[35] Directive 2018/1808, recital 14.
[36] C.T. Marsden, "Internet co-regulation and constitutionalism: Towards European judicial review International Review of Law Computers & Technology" 26(2):211-228, November 2012, https://www.researchgate.net/publication/254294662_Internet_co-regulation_and_constitutionalism_Towards_European_judicial_review (last accessed 10 June 2020).
[37] Martin Husovec, "Over-blocking: When is the EU legislator responsible," February 2021, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3784149 (last accessed 17 February, 2021)
[38] United Nations Human Rights Council, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue, 2017. https://www2.ohchr.org/english/bodies/hrcouncil/docs/17session/A.HRC.17.27_en.pdf (last accessed 23 September 2020)
[39] David Kaye, "A New Constitution for Content Moderation," medium.com, 15 June, 2019. https://onezero.medium.com/a-new-constitution-for-content-moderation-6249af611bdf (last accessed 23 September 2020

- the Council of Europe Committee of Ministers,[40] [41] including research that it has commissioned.[42]
- innumerable academic institutes such as the Amsterdam Institute for Information Law,[43] FGV Direito Rio,[44] and the Stanford University Center for Internet and Society.[45]
- the work coordinated and/or carried out by NGOs such as AccessNow,[46] Article 19,[47] Electronic Frontier Foundation,[48] epicenter.works,[49] European Digital Rights,[50] and Meedan,[51]

The list of outstanding work in this field is far too long to list every example.

Issues of relevance to content moderation have also been researched by the Internet & Jurisdiction project, in a broad way in relation to jurisdictional issues, most recently in its Global Status Report 2019[52] and in the form of two sets of specific, granular recommendations on "Mappings of Key Elements of Content Moderation."[53]

Instead of redrafting or re-imagining the comprehensive body of work, the Guidance Note and this Explanatory Memorandum take a step back and look at the framework for content moderation, in order to establish high-level guidelines for States on building approaches to content moderation that are both human rights-compatible and that achieve their public policy objectives, as well as to guide private companies.

The Guidance Note looks at broader issues, like how to develop a better understanding of the nature of the specific problems that content moderation seeks to address, how to ensure appropriate accountability for restrictions on human rights, the concepts of self- and co-regulation, and the characteristics of transparency tools that are fundamental to ensuring that goals are set and achieved.

---

[40] Recommendation CM/Rec(2018)2of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries. https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680790e14 (last accessed 13 October 2020).

[41] Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights implications of algorithmic systems. https://search.coe.int/cm/pages/result_details.aspx?objectid=09000016809e1154 (last accessed 13 October 2020).

[42] Such as Alexander Brown, "Models of Governance of Hate Speech Online," May 2020, https://rm.coe.int/models-of-governance-of-online-hate-speech/16809e671d (last accessed 15 January 2021).

[43] https://www.ivir.nl/publications/technology-and-law/.

[44] Luca Beli et al, Platform Regulations: "How platforms are regulated and how they regulate us," December 2017, https://direitorio.fgv.br/publicacoes/platform-regulations-how-platforms-are-regulated-and-how-they-regulate-us (last accessed 19 February, 2021).

[45] https://cyberlaw.stanford.edu/focus-areas/intermediary-liability

46 AccessNow, "26 recommendations on content governance," 2020. https://www.accessnow.org/cms/assets/uploads/2020/03/Recommendations-On-Content-Governance-digital.pdf (last accessed 23 September, 2020).

[47] Article 19, "Side-stepping Rights: Regulating Speech by Contract," 2018, https://www.article19.org/wp-content/uploads/2018/06/Regulating-speech-by-contract-WEB-v2.pdf (last accessed 4 January 2020)

[48] See santaclaraprinciples.org (last accessed 23 September, 2020).

[49] See platformregulation.eu (last accessed 23 September, 2020).

50 See https://edri.org/?s=content+moderation(last accessed 23 September, 2020).

[51] Meedan, "Content Moderation Toolkit, November 2018, https://meedan.com/reports/content-moderation-toolkit/ (last accessed 19 February 2021).

[52] Dan Jerker B. Svantesson, "Internet and Jurisdiction Global Situation Report 2019," https://www.internetjurisdiction.net/uploads/pdfs/GSR2019/Internet-Jurisdiction-Global-Status-Report-2019_web.pdf (last accessed 28 September, 2020).

[53] Internet & Jurisdiction Project, "I&J Outcomes: Mappings of Key Elements of Content Moderation," June 2020, https://www.internetjurisdiction.net/news/i-j-outcomes-mappings-of-key-elements-of-content-moderation (last accessed 28 September, 2020).

Understanding the problems

Content moderation is a tool used to address a wide variety of different problems. It is an element in the fight against serious crime online, against other online offences, against content that may be prejudicial to some audiences and against content that may be problematic for the business model of online companies (off-topic content on a specialised platform, for example).

Once one delves into the peculiarities of these challenges, one sees that a one-size-fits-all solution may often be possible but is rarely desirable. The consequences of a platform simply deleting (or not deleting) an off-topic post in a specialised forum are radically different from a platform simply deleting a video of a serious crime being committed, for example.

Regardless of the problem being addressed by content moderation, removal of an online post is a limitation of a user's freedom of expression, so this also needs to be done in a way which is predictable, legitimate, necessary and proportionate.

States should also not assume that internet intermediaries are either best placed to make decisions on legality or illegality of content or are neutral when making such decisions. When making a decision about whether to leave a piece of content online, a private company would be ill-suited to balance the rights of the complainant with the rights of the person uploading the content. It is even less suited to balance rights when its own interests are in play.[54] Twitter's share price fell over 10% after it permanently suspended the account of then US President Donald Trump, due to the expected impact that this would have on engagement with content on the platform.[55]

Whose responsibility?

When considering responsibility for such restrictions, we need to consider the fact that no content moderation is perfect and that they can be imposed as private decisions of intermediaries, decisions directly attributable to state regulation or a mix of the two.

"The unfathomable scale of online speech makes enforcement of rules only ever a matter of probability: content moderation will always involve error, and so the pertinent question is what error rates are reasonable and which kinds of errors should be preferred."[56] This requires clarity regarding responsibility for setting those targets, for the ensuing restrictions, and transparency permitting these decisions to be meaningfully audited.

Traditionally, internet intermediaries have preferred not to publish meaningful data on their own decisions, although they have been transparent on decisions made by states and others that have been imposed on them. This reflects a self-interest in not attracting scrutiny of their decisions. As is generally true in this policy area, it cannot be expected that internet intermediaries will voluntarily act against their own perceived self-interest and, therefore, specific legal obligations on transparency and methodology are necessary.

[54] For a detailed analysis of issues surrounding "balancing of rights" see, Evelyn Douek, "Governing Online Speech: From 'posts-as-trumps' to proportionality and probability" Columbia Law Review, Vol. 121, No. 1, 2021. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3679607 (last accessed 15 January, 2021)

[55] Business Insider, "Twitter loses $5 billion in market value after Trump is permanently banned from the platform," 11 January 2020, https://markets.businessinsider.com/news/stocks/twitter-stock-price-president-donald-trump-permanently-banned-tweeting-2021-1-1029946778 (last accessed 11 January, 2020)

[56] Douek, Evelyn, 2021, op cit, p. 1.

"Self-regulation" is understood differently in different contexts

We are used to self-regulation in the traditional media environment. Media companies establish their rules for the quality of their output and editorial decisions, as a mechanism for maintaining their independence and quality. They literally regulate themselves.

The situation is much more complicated in relation to content moderation. It can be an entirely internal process regarding, for example, how quickly complaints about content are processed, which fully fits the notion of "self-"regulation. However, decisions to remove user content is a regulation of users' speech, so not literally regulation of "self," but regulation of users and, indirectly, regulation of those who would otherwise have encountered that content. This raises different, and also very serious, considerations for democracy and human rights.

In addition, nominally self-regulatory initiatives can be undertaken in cooperation with, and/or with encouragement from governments, which makes it a cooperative endeavour between industry and governments, or more akin to "co-regulation". The extent to which the content moderation incurs the human rights responsibility of the state depends heavily on where on the continuum between self- and co-regulation the specific activity falls. States may also be held responsible if they do not take action or if no regulation is in place to prevent/remedy violations.

The scale of the activity is also entirely different compared with traditional media self-regulation. A 24-hour TV channel generates a specific and predictable amount of video every day, while 24 hours of video is uploaded to YouTube every three seconds.[57]

Success or failure?

Content moderation is a tool. It is used to achieve a goal or, as we have seen, a whole range of goals using a whole range of structures, such as self- and co-regulation to achieve those goals. But what if it does not work? What if the goals are never clearly defined? What if it works, but the cost is too high? What if it starts working and then stops working? What if it does not have a clear target? These questions, and the answers to them, can only be addressed if they are clearly articulated and if the data is collected to enable them to be answered.

II Problem statement

This section looks at key questions raised by content moderation.

The section starts by looking at the fact that a wide range of different types of content can be subject to content regulation (ranging from explicit legal prohibitions to moderation of legal content), with varying implications for human rights.

The section then looks at how the internal rules of internet companies are developed and enforced, with particular attention to standardisation and enforcement of both the rules, and the technologies used to enforce those rules.

Finally, the section looks at the complex balance of the roles and responsibilities of states and private companies, in an environment where private parties often impose restrictions as a result of direct or

[57] Statista, "Hours of video uploaded to YouTube every minute as of May 2019," https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/ (last accessed 13 October, 2020).

indirect government pressure and/or due to liability for failing to implement laws, which are sometimes unclear. It carries out this analysis from five perspectives:

- the implications of the privatisation of State policies on content restriction
- the challenge of ensuring the predictability of content restrictions of all kinds that are implemented via enforcement of private terms of service;
- the enduring lack of clarity regarding the balance between private ordering and state regulation in relation to, for example, enforcement of unclear or badly formulated law or procedures that directly or indirectly impact on content moderation;
- restrictions imposed for public policy reasons without these being prescribed by law;
- the need to choose metrics carefully, to ensure that bad practice (such as excessive focus on the speed or volume of content removed) can be avoided.

1. Prevalence of unwelcome/abusive content/behaviour

It is crucial to recognise that different subjects of content moderation are fundamentally different problems and, therefore, that "one size fits all" solutions may not be appropriate. Chapter IV details six broad categories of illegal or potentially problematic content, ranging from content that is illegal everywhere to content that is legal but potentially harmful. It is worth noting that, with the partial exception of child abuse images, there is little harmonised legislation internationally on any of these types of content. This raises significant jurisdictional issues where the uploader could be in one country, the downloader in a second country and the service provider in a third country.[58] This increases legal uncertainty for intermediaries and, therefore, increases risks of, for example, over-blocking, extra-territorial implementation of national laws, etc.[59]

It is therefore important for states to design a structured methodology for responding rapidly, proportionately, and effectively to significant problems. This is essential, but currently absent. For example, policy responses to the spread of conspiracy theories on topics as diverse as 5G mobile communications and vaccines have so far been generally slow and inadequate.[60] Such content can have significant real-world consequences, ranging from the burning of mobile communication masts to outbreaks of diseases that were previously under control. Another example is the feared rise of "deep fakes." This technology permits existing videos to be convincingly recreated, with a visual or audio component (for example a person) in the video being replaced by somebody/something else. This is an example of content that is not necessarily part of a wider offence and where the content itself is not necessarily illegal. It is better to have strategies in place that can be used if such phenomena become a problem, rather than reacting to them when they do.

2. Lack of standardisation of content restrictions

Under self- and co-regulatory schemes, restrictions are usually imposed based on the internal rules of the internet intermediary. The naming and number of these internal rules vary from company to company.[61]

---

[58] For an introduction to this topic see: Graham Smith, "Peaceful coexistence, jurisdiction and the internet," 25 February 2018, https://www.cyberleagle.com/2018/02/peaceful-coexistence-jurisdiction-and.html (last accessed 1 September 2020).

[59] See Dan Jerker B Svandesson, 2019, op cit.

[60] See, for example, EU Disinfo Lab, "Covid-19 and 5G: A Case Study of Platforms' Content Moderation of Conspiracy Theories," 14 April, 2020, https://www.disinfo.eu/publications/coronavirus-and-5g-a-case-study-of-platforms-content-moderation-of-conspiracy-theories, (last accessed 04 January, 2021).

[61] Facebook, for example, has "terms of service" (https://www.facebook.com/legal/terms) and "community standards" (https://facebook.com/communitystandards) for its social media service users, while Twitter has "Twitter rules." (https://help.twitter.com/en/rules-and-policies/twitter-rules) (all last accessed 08 October 2020)

It is symptomatic of the complexity of the issues at hand that we suffer both from a lack of standardisation (regarding how individual companies interpret and apply their internal rules) and from too much unilateral standardisation by the biggest providers.

We suffer from too little standardisation insofar as the meanings of the words in internet intermediary terms of service are often unclear. Even more problematic is the fact that the terms used are also subject to re-interpretation.

For example, Propublica found that Facebook's content moderators came to different conclusions on whether to delete broadly similar content and did not always abide by company guidelines on how to treat content.[62] In essence, broadly similar content is, and is not, banned by Facebook, and its mechanisms for enforcing its internal rules are not always respected.

This problem is exacerbated by terms of service being written to be vague, possibly deliberately, in order to allow intermediaries maximum flexibility to act if they are put under pressure by, for example, states.[63] A particularly clear illustration of the flexible meaning of Facebook's agreements with its users can be found in its 2012 unannounced experiment to establish if the company could manipulate the mood of approximately 70,000 of its users. Faced with media criticism, Facebook explained that users had signed up to be researched upon in this way because "when someone signs up for Facebook, we've always asked permission to use their information to provide and enhance the services we offer. To suggest we conducted any corporate research without permission is complete fiction."[64] Oddly, however, the company did add "research" to its terms of service a few months later, despite having previously held that this was already clearly a use to which personal data could be put. We suffer from too much standardisation in that key aspects of terms of service and use of specific tools for filtering by the largest internet intermediaries create de facto global standards with a potentially significant impact on human rights, without multi-stakeholder or democratically legitimated decision-making. So, while terms of service are frequently vague and unpredictable, the content restrictions that are based on them are being harmonised/standardised by global intermediaries. This is a two-edged sword of course, the more room for arbitrariness they accord themselves, the more this arbitrariness can be exploited by external pressure.

The rules and technologies used to implement them at scale are becoming more homogenous, as explained by Evelyn Douek in her essay "The Rise of the Content Cartels."[65] Ms Douek describes the processes that have led to a very small number of major internet companies setting the standard for what is permitted online globally, as well as the technologies used to implement this standard, to the detriment of transparency and accountability. This happens through nominally self-regulatory initiatives such as the Global Internet Forum to Counter Terrorism (GIFCT).[66] Such initiatives normally start with involvement of the key global companies and certain States, which set the rules and define the technologies and filtering lists to be implemented. They are then joined, often under government

[62] Ariana Tobin, Madeleine Varner and Julia Angwin, "Facebook's Uneven Enforcement of Hate Speech Rules Allows Vile Posts to Stay Up," Propublica, 28 December 2017, https://www.propublica.org/article/facebook-enforcement-hate-speech-rules-mistakes (last accessed 05 May 2020).

[63] Article 4(1)m of Regulation 2016/794 of the European Parliament and of the Council, 11 May 2016, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0794 (last accessed 07 May 2020).

[64] Alex Hern, "Facebook T&Cs introduced 'research' policy months after emotion study," The Guardian, 01 July 2014, https://www.theguardian.com/technology/2014/jul/01/facebook-data-policy-research-emotion-study (last accessed 05 May, 2020).

[65] Evelyn Doudek, "The Rise of Content Cartels", Knight First Amendment Institute, February 2020, https://knightcolumbia.org/content/the-rise-of-content-cartels (last accessed 05 May 2020).

[66] See gifct.org (last accessed 20 May 2020).

pressure by smaller companies, with little control over or input into, the restrictions they are "voluntarily" imposing.[67]

Despite not respecting principles of multistakeholderism,[68] such standardisation does not necessarily undermine human rights. To ensure human rights are protected, states should ensure that three core conditions are met: a) there is clear agreement on the illegality of the content that is being curtailed, in full compliance with human rights law, b) there is transparency regarding engagement of law enforcement authorities in cases where evidence of such crimes is detected and, c) there is rigorous transparency regarding the content being removed, the criteria for which are described in Section V below. In order to have effective and lawful restriction of illegal content, we need clear rules, with maximum transparency. Where rules are unclear and/or subject to regular and poorly communicated changes, designed in an opaque manner and imposed arbitrarily, neither the scope of the rules nor the sanctions for breaking them are known. Furthermore, it is not clear what the impact on the crime(s) being addressed may be.

So, at one level we have unclear terms of service being inconsistently implemented by individual internet intermediaries and, on another, we have global internet giants setting standards for both what is filtered and how.

Furthermore, with regard to content that is in a grey area, such as content that may have the "intent" of "incitement to terrorism," the law can offer little assistance for an intermediary to be able to assess illegality, a role which an interested party is ill-suited to fill in any case. When such content is removed, there is currently no transparency regarding, for example, the involvement of law enforcement authorities or independent ex post assessment of whether the restricted content was, in fact, illegal. Where the technologies that may be used are intrusive or disproportionate, such as the blocking of certain words, phrases and images, or using technology to guess the intent behind words, phrases or images, the predictability of content restrictions becomes even weaker.

> "*Current content cartels, as well as the future ones we are likely hurtling toward, allow participants to launder difficult decisions through opaque processes to make them appear more legitimate than they really are and do not mitigate the threat of a handful of actors holding too much power over the public sphere*."[69]

All measures undertaken to restrict content are restrictions of freedom of expression. States have positive and negative obligations to ensure that restrictions under regulatory, self- and co-regulatory regimes are legally acceptable. To be acceptable, they must be predictable, legitimate, necessary, and proportionate. Lack of clarity and lack of independent supervision mean that these obligations are not currently being achieved.

This raises critical issues that policymakers need to consider:

---

[67] The "European Commission Recommendation on measures to effectively tackle illegal content online" stresses the need to engage with smaller operators to assist in roll-out of content restriction technologies. For example, recital 37 which recommends roll-out of proactive content measures to companies that do not have the scale to operate them. "Those cooperative efforts are particularly important to help enabling hosting service providers which, because of their size or the scale on which they operate, have limited resources and expertise to respond effectively and urgently to referrals and to take proactive measures, as recommended." Recommendation (2018)1177 on measures to effectively tackle illegal content online," March 2018, https://ec.europa.eu/digital-single-market/en/news/commission-recommendation-measures-effectively-tackle-illegal-content-online (last accessed 14 October 2020).

[68] Multistakeholderism is a widely promoted principle of internet governance, whose purpose is to ensure that all stakeholder interests are adequately taken into consideration in the design and implementation of solutions to common problems or to achieve common goals.

[69] Evely Doudek "The Rise of the Content Cartels" 2020, op cit.

- Firstly, policymakers should take account of the fact that content moderation addresses a wide variety of different kinds of unwelcome content and that a one-size-fits-all approach is unlikely to be effective or proportionate.
- Secondly, the restrictions imposed under content moderation are usually based on the internal rules of internet intermediaries, which are often unclear and implemented in an unpredictable way, falling below the requirements of the European Convention on Human Rights.

Finally, while international, coordinated action can lead to human rights-friendly and effective measures, such coordination needs to implement human rights by design and must fully engage all stakeholder groups in both the design and implementation phases.

3. The line between public and private

The European Convention on Human Rights and the case law of the European Court of Human Rights provide an extraordinarily rich and living framework within which human rights can be nurtured and protected. The Convention is binding on States, in the form of both positive and negative obligations for the protection of those rights. The rights in the Convention are the default, with restrictions being exceptions that must be justified.

Governments therefore have two tasks regarding content moderation online:

- to fulfil their positive and negative obligations under the Convention and
- to ensure the enforcement of national law.

Properly respected, the two tasks generate a great deal of synergy between the obligations to respect the Convention and to ensure enforcement of national law. . If states rely on private companies and engage in co-regulatory schemes, those states have to ensure that the principles or conditions of restrictions are followed (such as in relation to predictability and proportionality).[70] It has been argued that human rights safeguards improve the quality of the law.[71]

A co-regulatory project that seeks to ensure predictability and proportionality will need tools to assess the initial and ongoing efficiency and proportionality of the measure. This, in turn, will drive better and more targeted enforcement of the law and allow policymakers to track the evolution of the problem being addressed. A self- or co-regulatory project that does not have targets, options for adaptation or abandonment, or independent supervisory mechanisms will neither be able to demonstrate proportionality nor effectiveness.

In his report of 16 May 2011, the UN Special Rapporteur on Freedom of Opinion and Expression stated that "censorship measures should never be delegated to a private entity".[72] He gave one example of a "quasi-State and quasi-private entity tasked to regulate online content" as an example of this kind of delegation of power. Such delegation does sometimes happen through direct or indirect pressure by states on internet intermediaries. Nonetheless, many internet intermediaries can and do

[70] See Costello-Roberts vs UK, application 13134/87 (para 27) "the State cannot absolve itself from responsibility by delegating its obligations to private bodies or individuals" cited in Kuczerawy, A, "The power of positive thinking: intermediary liability and the effective enjoyment of the right to freedom of expression,", August 2017, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3033799 (last accessed 5 January, 2021).

[71] M. Husovec, op cit, p.12

[72] Human Rights Council of the United Nations, "Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue," 16 May, 2011, https://www2.ohchr.org/english/bodies/hrcouncil/docs/17session/A.HRC.17.27_en.pdf (last accessed 06 May 2020). See also Human Rights Council of the United Nations, "Report of the Special Rapporteur of the promotion and protection of the right to freedom of opinion and expression, David Kaye, 06 April 2018, https://www.undocs.org/A/HRC/38/35 (last accessed 04 January 2020).

legitimately restrict content as part of their contractual freedom. A voluntary restriction, such as limiting or removing clearly defined kinds of content on an online platform aimed at children or at a particular profession (a platform for discussion of medical issues, for example) would not normally pose problems for freedom of expression.

The dividing line between legitimate co-regulation, illegitimate state pressure on intermediaries and legitimate enforcement by intermediaries of their internal rules has been notoriously difficult to draw. The risk of privatised censorship is particularly grave in situations where states have constructed a liability regime that incentivises nominally contractual restrictions by intermediaries and this can be further aggravated if the law that is being implemented is, itself, not clear. States, individually and collectively, have a positive obligation to mitigate this risk by ensuring a clear and predictable legal framework, where both the laws prohibiting certain types of content and rules on liability are clear.

There is considerable scope for ill-defined, potentially counter-productive (both from the perspective of human rights and the public policy objectives being addressed) restrictions being imposed by internet intermediaries, under direct (such as by demanding a code of conduct) or indirect (such as by implementing or even having excessive liability rules) government pressure. The UN Special Rapporteur gave a clear example of how the actions of states can lead to restrictions on human rights that amount to "censorship by proxy"[73]:

> *"However, while a notice-and-takedown system is one way to prevent intermediaries from actively engaging in or encouraging unlawful behaviour on their services, it is subject to abuse by both State and private actors. Users who are notified by the service provider that their content has been flagged as unlawful often have little recourse or few resources to challenge the takedown. Moreover, given that intermediaries may still be held financially or in some cases criminally liable if they do not remove content upon receipt of notification by users regarding unlawful content, they are incentivised to err on the side of safety by over-censoring potentially illegal content. Lack of transparency in the intermediaries' decision-making process also often obscures discriminatory practices or political pressure affecting the companies' decisions. Furthermore, intermediaries, as private entities, are not best placed to make the determination of whether a particular content is illegal, which requires careful balancing of competing interests and consideration of defences*."[74]

[73] Aleksandra Kuczerawy, "Private enforcement of public policy: freedom of expression in the era of online gatekeeping," Liras, June 2018.
[74] Human Rights Council of the United Nations 2011, op cit, para. 42.

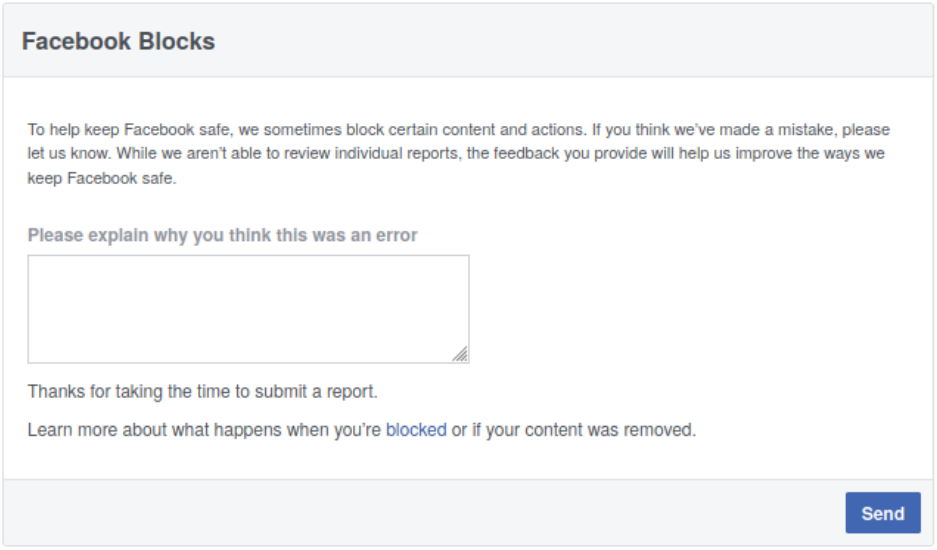

Figure 1: Message received by a Facebook user in response to a complaint that the site blocks users from posting links to his entirely legal website.

The scale of non-law-based restrictions imposed by private companies for ostensibly public policy reasons was also made clear in a study undertaken by the Swiss Institute for Comparative Law for the Council of Europe.[75] That research identified widespread blocking of online resources by internet access providers without a clear legal basis under national law.[76] This blocking was done on the basis of so-called self-regulatory initiatives and frequently involved online resources where the content itself was not illegal (alleged copyright infringements) or where it amounted to evidence of criminal behaviour (child abuse material). In the latter case, the blocking was being done without treating this information with the seriousness it deserved, bearing in mind it was alleged to be evidence of a serious crime.

The more urgent it is to engage in such restrictions, such as in relation to child abuse material, the more important it is to ensure accountability. Implementation of effective transparency, review and adjustment mechanisms are crucial to ensure effectiveness and proportionality. Sadly, this is not always the case. In a Resolution adopted by a huge majority (597 in favour, 6 against[77]) in 2017, the European Parliament made this very clear in relation to the fight against child abuse material online:

> *"whereas the Commission's implementation report does not provide any statistics on the take-down and blocking of websites containing or disseminating images of child sexual abuse, especially statistics on the speed of removal of content, the frequency with which reports are followed up by law enforcement authorities, the delays in take-downs due to the need to avoid interference with ongoing investigations, or the frequency with which any such stored data is actually used by judicial or law enforcement authorities*;"[78]

[75] Swiss Institute for Comparative Law, "Comparative study on blocking, filtering and take-down of illegal internet content," Council of Europe, 2017. https://edoc.coe.int/en/internet/7289-pdf-comparative-study-on-blocking-filtering-and-take-down-of-illegal-internet-content-.html (last accessed 4 May 2020).

[76] This is not, strictly speaking, "content moderation" in the usual sense. Nonetheless, it is worthy of note in this context due to the many overlapping characteristics with content moderation.

[77] See https://oeil.secure.europarl.europa.eu/oeil/popups/sda.do?id=30474&l=en (last accessed 12 January 2021).

[78] European Parliament Resolution of 14 December 2017 on the implementation of Directive 2011/93/EU of the European Parliament and of the Council of 13 December 2011 on combating the sexual abuse and sexual exploitation of children and

a) Unpredictability of content restrictions

Implementation of restrictions on the basis of the internal rules of an internet intermediary limit the freedom of expression of individuals. The European Convention on Human Rights requires restrictions to its Articles 8-11 to be provided for by law. This does not mean that restrictions cannot be imposed by internet companies, but it does mean that such rules should be clear and applied consistently so that users can adapt their behaviour.[79] This is particularly true of restrictions implemented with the cooperation of the State under co-regulatory procedures, where the levels of clarity would need to respect the requirements for predictability provided for in the Convention and relevant case law and should not amount to improper delegation of censorship powers to private parties.

As we saw above, in the case of self- and co-regulatory approaches, contractual documents (variously called "terms of service," "community guidelines," etc) define the limits of what is or is not permitted and are generally applied in the framework of self- and co-regulatory restrictions.

A study undertaken in 2017 for the Council of Europe by the FGV Direito Rio looked at the issue of human rights and online platform contracts.[80] It describes terms of service as follows:

> *"Terms of Service are standardized contracts, defined unilaterally and offered indiscriminately on equal terms to any user. Since users do not have the choice to negotiate, but only accept or reject these terms, Terms of Service are part of the legal category of adhesion agreements. In fact, these agreements establish a kind of 'take it or leave it' relationship, replacing the traditional concept of bargained clauses among contracting parties*."[81]

The research undertaken for that study found that "such terms are generally long, dense and formulated in language that is hard to be understood by anyone who does not have legal training […] people hardly ever read these contracts […]. When they do, they find them difficult to understand."[82]

As an example of what this means for the predictability of content restrictions, the study found that, out of 50 service providers, 23 had either contradictory (20) or no (3) provisions in their terms of service as to whether they analyse, block or remove content for "somewhat specific, undetermined or unclear reasons".[83]

It is clear, therefore, that predictability of the meaning and enforcement of the internal rules of internet companies frequently fall short of the standards needed to ensure the protection of human rights. Failures should be identified and rectified before any self- or co-regulatory scheme that relies on such rules is launched.

---

child pornography, https://www.europarl.europa.eu/doceo/document/TA-8-2017-0501_EN.html (last accessed 27 May, 2020)

[79] "Thus, a norm cannot be regarded as a 'law' unless it is formulated with sufficient precision to enable citizens to regulate their conduct; they must be able –if need be with appropriate advice –to foresee, to a degree that is reasonable in the circumstances, the consequences which a given action may entail." Application number 38433/09, Centro Europa 7 S.R.L. and Di Stefano v.Italy, 2012, paragraph 141 http://hudoc.echr.coe.int/eng?i=001-111399 (last accessed 14 October 2020)

[80] Venturini, J., Louzada, L., Maciel, M.F,, Zingales, N., Stylianou, K., Belli, L, "Terms of service and human rights: an analysis of online platform contracts," Editora Revan, 2016, https://internet-governance.fgv.br/sites/internet-governance.fgv.br/files/publicacoes/terms_of_services_06_12_2016.pdf (last accessed 08 June 2020)

[81] Venturini et al, op cit, 2016, p.23, citing Lemley, M. A. (2006). *Terms of Use*, 91 MINN. L. REV. 459, 459 http://www.kentlaw.edu/faculty/rwarner/classes/ecommerce/2008/contracts/consent/lemley%20tersm%20of%20use.pdf

[82] Ibidp.24

[83] Ibid, p.54

b) Lack of clarity on balance of roles & responsibilities of states and private actors

These considerations create a highly complex set of criteria for assessing the fulfilment of positive and negative state obligations in relation to restrictions of human rights through self- and co-regulatory measures.

As the UN Special Rapporteur pointed out, even something as superficially uncontroversial as a "notice and takedown" system, such as is used widely in the Council of Europe region, can fail to provide sufficient protection for freedom of expression.[84]

Restrictions are sometimes implemented as a result of unclear liability laws and sometimes as a means of working around the problem that laws on illegality of content are, themselves, unclear. For example, a project funded by the European Commission found that there were "huge disparities" in the EU between national laws forbidding racism and xenophobia despite their being based on EU legislation.[85]

Faced with disparity of national laws in Europe, even the European police agency Europol is legally obliged to report online content that is potentially linked to serious crime as possible terms of service violations "for the voluntary consideration" of internet intermediaries, rather than a possible breach of the law.[86] The possibly (or possibly not) illegal content in question is apparently serious enough to warrant action from Europol, but not serious enough for the companies to be required to take action or to be definitively outlawed under EU law.

The very unclear situation in which internet intermediaries find themselves was laid bare by a response to a Parliamentary Question on this topic to the European Commission.[87] As loss of immunity from liability is triggered when the intermediary gains "actual knowledge" of illegal content, a Parliamentarian asked if a referral from Europol would constitute such knowledge. The European Commission responded that it would not constitute "actual knowledge" unless the provider was aware "of fact [sic] and circumstances on the basis of which a diligent economic operator should have identified the illegality in question." In other words, even a referral from Europol would not necessarily be enough to trigger knowledge of illegality, even if it is possible that, in some undefined[88] circumstances, it might be.[89] It is unclear why a law enforcement authority would not be required in such circumstances to take appropriate steps in cooperation with judicial authorities and national law

[84] The United Nations (UN) Special Rapporteur on Freedom of Opinion and Expression, the Organization for Security and Co-operation in Europe (OSCE) Representative on Freedom of the Media, the Organization of American States (OAS) Special Rapporteur on Freedom of Expression and the African Commission on Human and Peoples' Rights (ACHPR) Special Rapporteur on Freedom of Expression and Access to Information, "Joint Declaration on Freedom of Expression and the Internet," June 2011https://www.oas.org/en/iachr/expression/showarticle.asp?artID=848 (last accessed 04 January 2020), paragraph 2b.

[85] Mandola Project, "Definition of illegal hatred and implications," 31 March 2016, page 7, http://mandola-project.eu/m/filer_public/7b/8f/7b8f3f88-2270-47ed-8791-8fbfb320b755/mandola-d21.pdf (last accessed 15 May 2020).

[86] Article 4(1)m of Regulation 2016/794 of the European Parliament and of the Council, 11 May 2016, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0794 (last accessed 07 May 2020).

[87] Question E-7205/17 by Cornelia Ernst to the European Commission, 23 November 2017. https://www.europarl.europa.eu/doceo/document/E-8-2017-007205_EN.html (last accessed 05 January 2021)

[88] The wording in the European Commission's response comes from a case from the European Court of Justice related to the unauthorised sale of cosmetics and not illegal content Case 324/09, L'Oréal SA and Others v eBay International AG and Others. http://curia.europa.eu/juris/document/document.jsf?text=&docid=107261&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=11354812 (last accessed 27 July 2020).

[89] Parliamentary question E-7205/2017, https://www.europarl.europa.eu/doceo/document/E-8-2017-007205_EN.html (last accessed 20 May, 2020).

enforcement bodies rather than having the obligation to rely on an internet intermediary's terms of service.

In a situation where the internet intermediary receives a referral from Europol, there are several factors incentivising the former to remove the content:

- the fact that the referral came from a law enforcement agency (albeit without an explicit notification that the content was illegal);
- the possible reputational harms from not responding to such referrals by deleting the content in question;
- the fact that the referral may (or may not) be interpreted by a court as "actual knowledge" or "knowledge" of illegality, thereby creating liability for the intermediary, even though Europol was unable or unwilling to provide – via a referral to national judicial authorities, for example – confirmation of the illegality of the content.

There is little, however, beyond the company's business relationship with an individual customer, incentivising them to keep the content online.

In such an environment, who is at fault if operators regularly remove, as they are incentivised to do, legal content, without this being clearly predictable, necessary, or proportionate?

There are two possibilities. Firstly, that the decision is a private one, based on private rules which, when given the option to take it or leave it, individuals agreed to. In this case the action possibly falls outside the scope of human rights law and is the combined responsibility of the intermediary and the user. This argument appears untenable in light of European Court of Justice case law. For example, *Cengiz and Others v. Turkey*, where the internet was described as "one of the principal means by which individuals exercise their right to freedom to receive and impart information and ideas, providing as it does essential tools for participation in activities and discussions concerning political issues and issues of general interest".[90]

Secondly, that States

- failed individually and collectively to require terms of service to be clear;
- have not required terms of service to be implemented in a predictable and proportionate manner;
- omitted to adopt clear laws on either a national or international level regarding the illegal content in question;
- did not ensure that a competent authority was in place to issue an order to remove or not remove the content in question and
- did not provide a legal framework that ensured a more appropriate balance of incentives for providers.

If the latter is the case, then this falls short of the States' positive obligations under the European Convention on Human Rights and is in breach of the requirement that any restrictions must be provided for by law.

It is clear, however, that bigger internet intermediaries are better equipped to navigate such legal uncertainty than their smaller competitors. Similarly, bigger intermediaries are better placed to cope with more onerous liability rules (such as through buying or developing filtering technologies to comply with very short deadlines for removing content) than their smaller competitors. At a time

[90] Cengiz and Others v. Turkey, application numbers 48226/10 and 14027/11, para 49.

when policymakers increasingly voice concerns about the growing power of the largest platforms, this should be a significant consideration for States.[91] Consequently, for the sake of human rights and also to ensure innovation and competition, unclear or onerous liability rules should be rigorously avoided.

In conclusion, it is incumbent on the State to ensure that the liability rules for internet intermediaries are sufficiently clear to avoid incentivising privatised censorship and that the internal rules of internet intermediaries, and their implementation, are clear, predictable, and proportionate. Otherwise, States are creating incentives for restrictions on freedom of expression that fail to clearly respect the safeguards for permissible restrictions laid down in the Convention *and* allowing this to happen in an environment where individuals under their protection are defenceless against the vagaries of profit-motivated internet intermediaries.

c) Dangers of over-compliance, especially when leading to discriminatory outcomes

Overcompliance:

In an environment where an internet intermediary can be held liable when failing to remove content or services that might constitute an infringement, but has few counterbalancing incentives to keep content online, it seems inevitable that intermediaries will restrict content that falls in the "grey zone," relying on their terms of service to do so.[92] As a result, ideas, explicitly protected by the European Convention on Human Rights, such as those "that offend, shock or disturb the State or any sector of the population,"[93] have little chance of being protected in practice. Research indicates that "providers tend to over-remove content to avoid liability and save resources, they equally employ technology to evaluate the notifications; and the affected users who posted content often do not take action".[94]

Discrimination:

Moreover, issues of discrimination and gaming (i.e., deliberate, not necessarily illegal, manipulation or abuse of intermediaries' complaints systems) arise in a system that relies heavily on self- and co-regulatory approaches to address unwelcome or illegal content online and that lacks a clear legal framework for these restrictions.

In the absence of an obligation to do so, there is no reason why, accidentally or deliberately, an operator might not prioritise dealing with racist abuse against one group more than another, or one form of sexism over another.[95] Indeed, in the absence of clear rules on transparency, particularly when using artificial intelligence, there would be no way of knowing whether the discriminatory approach was even happening. It seems counterintuitive that there is increasing reliance on artificial intelligence to make sometimes highly complex decisions about the lawfulness, terms of service compliance and

---

[91] Mark Sweney, "Google and Facebook dominance should be curbed, suggests CMA," The Guardian, 18 December 2019, https://www.theguardian.com/business/2019/dec/18/google-facebook-dominance-curbed-cma-report-uk-digital-market (last accessed 9 July 2020).

[92] Lack of transparency means that there is little empirical data on this. However, there is such a large number of examples that this phenomenon seems difficult to deny. See, for example New York Times, "YouTube is erasing history," 23 October 2019, https://www.nytimes.com/2019/10/23/opinion/syria-youtube-content-moderation.html (last accessed 18 January 2020)

[93] *Handyside v. the United Kingdom*, no. 5493/72, 7 December 1976, para. 49. https://hudoc.echr.coe.int/app/conversion/pdf/?library=ECHR&id=001-57499&filename=001- (pdf)(last accessed 9 July 2020).

[94] Alexandre De Streel et al "Online Platforms' Moderation of Illegal Online Content". Study undertaken for the European Parliament's Internal Market and Consumer Protection Committee, June 2020, https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652718/IPOL_STU(2020)652718_EN.pdf (last accessed 18 January 2021)

[95] Maarten Sap et al, "The Risk of Racial Bias in Hate Speech Detection," Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1668–1678, Florence, Italy, July 28 - August 2, 2019.

context of speech and yet these sophisticated technologies rarely, if ever, produce timely, thorough reports regarding the specific reasons why content was removed (or not), nor granular detail regarding how much content was restricted for those reasons.

> *"Ascertaining whether an AI system impacts human rights, democracy and the rule of law can be rendered difficult or impossible when there is no transparency about whether a product or service uses an AI system, and if so, based on which criteria it operates. Further, without such information, a decision informed or taken by an AI system cannot be effectively contested, nor can the system be improved or fixed when causing harm*."[96]

Yet, there is considerable evidence of discriminatory "gaming" of complaints systems of online intermediaries.[97] Without the transparency needed to identify mistakes and to facilitate analysis of why mistakes happened when they did, the problems appear destined to continue. One particularly egregious example was the case of the women's rights group "Women on Waves." YouTube removed the organisation's entirely legal and terms of service compliant channel four times, without explanation, in the course of 2018.[98] Another example is the temporary removal of the anti-racism "Kick Out Zwarte Piet" page from Instagram while a page allegedly inciting violence against that movement remained online.[99]

Internet intermediaries may, not entirely without justification, argue that lack of transparency about how such decisions are made is necessary to prevent further gaming. However, using lack of transparency as an ostensible way of protecting the integrity of malfunctioning internal procedures externalises the cost of such (unproven) procedures to the victims of incorrect and inconsistent decisions, who are left with no way of avoiding similar decisions in future and no recourse.

States should take action where necessary against over-compliance and discrimination by implementing rules on transparency and by providing clear guidance or rules on the predictability and balance of terms of service agreements between intermediaries and their users as well as on the enforcement of those contracts.

d) Focus on metrics restricted to speed and volume, driven by repeated "urgent" situations

Technology, crime and society change continuously. As a result, it is inevitable that the environment in which a self- or co-regulatory measure is implemented will change in the short- or medium-term. It is therefore crucial to ensure that targets are met and adjustments can be made to any measure, as the situation changes.

All too often, when a particular topic hits the headlines, states and private actors need to be seen to be acting decisively. In such circumstances, it is easy for states to demand that internet intermediaries

---

[96] Council of Europe Ad Hoc Committee on Artificial Intelligence, "Feasibility study", December 2020 https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da (last accessed 21 January, 2021)

[97] Content from or about women seems to suffer disproportionately from this, ranging from bra manufacturer Thirdlove being banned from Facebook for displaying a box of bras (https://kernelmag.dailydot.com/issue-sections/features-issue-sections/12796/facebook-nudity-breasts-advertising/) to an advertisement about menstruation being blocked due to "excessive skin" (https://www.independent.co.uk/news/media/online/woman-gets-censored-by-facebook-because-she-blogs-about-periods-a6725176.html). See also the following footnote.

[98] Evelyn Austin, "Women on Waves' three YouTube suspensions this year show yet again that we can't let internet companies police our speech", Bits of Freedom, June 2018, https://www.bitsoffreedom.nl/2018/06/28/women-on-waves-three-youtube-suspensions-this-year-show-yet-again-that-we-cant-let-internet-companies-police-our-speech/ (last accessed 14 February 2020).

[99] Stan Hulsen, "Instagram-account van KOZP tijdelijk geschorst, haataccount nog wel online"(Instagram account from KOZP temporarily suspended, hate account remains online) Nu.nl, 20 November 2019, https://www.nu.nl/tech/6012319/instagram-account-van-kozp-tijdelijk-geschorst-haataccount-nog-wel-online.html (last accessed 22 May 2020) (in Dutch).

do "more" to fight the problem. Metrics, such as simple number of posts removed, for example, create a new incentive to delete "more" and "faster". Deleting content quickly is easy. However, this comes at unpredictable but unquestionable costs.

The first implementation report on the EU Code of Conduct on Countering Illegal Hate Speech Online provides a useful model to demonstrate some challenges, which are left to internet intermediaries to resolve.

The tasks that the code of conduct was meant to achieve were:

- to tackle "hate speech" as defined by the 2008 Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law[100] and prevent its spread, despite the "huge disparities" in the laws implementing that legislation[101] and,
- to defend freedom of expression.

From a positive perspective, the Code can be seen as a successful effort to move the debate outside policy discussions and into higher levels of the management structures of internet intermediaries.

However, it provides a useful case study regarding the pitfalls of self- and co-regulatory approaches driven by government pressure.[102]

Six months after the code was adopted, an implementation report entitled "Code of Conduct on Countering Illegal Hate Speech Online: First results of implementation" was published by the European Commission. [103]

Under one heading ("methodology"), it refers to the notifications being assessed as "notification of *alleged* illegal hate speech" (emphasis added). Under another heading ("illegal hate speech notifications made to IT companies") the same reports were referred to as "notifications of illegal hate speech" (i.e. the "allegations" are apparently all assumed to be valid) and this was repeated under subsequent headings. The only data provided in the monitoring exercise relate to types of allegedly illegal content, number of notifications, speed of processing of notifications, types of notifiers, and percentages of notifications that led to removals for each provider – and not the actual amount of correctly identified illegal content that was removed (through, for example, independent review of samples of content removed and left online, with agreed benchmarks for acceptable error rates).

If a significant proportion of the removed content was actually criminal and if the content that was not removed was actually legal, then these goals would be, at least in part, achieved. However, no data is generated in relation to these questions. We only learn about speed and quantity of removals. Even a minimal measure like random sampling of takedown and non-takedown decisions, that would give an indication of the current impact, would allow for adaptations to be made.

The impact of the measures on the crimes being committed has never been assessed. It could credibly be argued that efficient takedowns might dissuade some people from uploading criminal hate speech, meaning that the measure might be working. However, it could also be credibly argued that relying on the code of conduct creates a degree of impunity, as the worst possible sanction is an eventual

[100] Council Framework Decision 2008/913/JHA, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM%3Al33178 (last accessed 08 June, 2020).
[101] Mandola Project, 2016, op cit, p. 9.
[102] Government initiatives such as the German Network Enforcement Law suggest that states are becoming more aware of the limitations of this approach.
[103] European Commission, "Code of Conduct on countering illegal hate speech online: First results of implementation," December 2016, https://ec.europa.eu/newsroom/document.cfm?doc_id=40573 (last accessed 08 June 2020).

removal of posts that might otherwise have led to criminal prosecution, due to their illegality. Raw data on speed and volume of takedowns misses such nuances and fails to track the evolution of the problem over time.

In fact, according to the fifth monitoring report of the Code of Conduct, of the 3,099 pieces of possibly illegal content that were removed by the participating companies, 85% was not referred to any law enforcement authority by any participating organisation and no data is provided as to whether any enforcement action was taken in relation to the remaining 15%.[104]

e) Moderating risk

A further set of issues arise in relation to moderation of legal content to protect vulnerable groups, such as children. Importantly, "risk is not harm"[105] and the avoidance of risk may itself be harmful. Therefore, pressure on intermediaries to "self-regulate" to minimise risk to vulnerable groups brings with it its own set of dangers. For example, the UK Schools Inspectorate (OFSTED) found that "pupils were more vulnerable overall when schools used locked down systems because they were not given enough opportunities to learn how to assess and manage risk for themselves."[106] This does not mean that content moderation for the protection of vulnerable groups should not be done, but rather that it needs very careful oversight and review mechanisms. Creating a liability risk on the part of the internet intermediary, which needs to be mitigated by the intermediary eliminating or minimising the perceived risks to vulnerable groups (even if the potential benefits may, or are likely to, outweigh the potential harm), may not be the most effective, proportionate and targeted way towards achieving a balanced outcome.

As a result, in such circumstances, it is crucial for the avoidance of unintended consequences, that any policy interventions that have the purpose of minimising risk are clearly recognised as such, in order to mitigate the particular problems of this approach, with the state taking its share of responsibility. They should also have clear targets, adjustment mechanisms and supervision, meaningful protection for freedom of expression, as well as tools to identify counterproductive impacts.

III Affected rights

Self- and co-regulatory frameworks for content moderation foresee various measures to make the illegal activity more difficult. For example, they can refuse certain key words (e.g. "ivory") in an online advertisement in a jurisdiction where the sale of the product is illegal. The measures can also focus on the removal of content, activities or accounts that fall under the scope of the framework, based on complaints or notices received from third parties. YouTube (owned by Google), for example, globally removes the accounts of individuals who are subject to three correctly formulated copyright complaints under US law and Google search globally demotes domains that are subject to "large amounts" of correctly formulated complaints.[107]

---

[104] 5th Implementation of the Code of Conduct, European Commission, Factsheet June 2020, https://ec.europa.eu/info/sites/info/files/codeofconduct_2020_factsheet_12.pdf (last accessed 24 July, 2020).

[105] Livingstone, Sonia; Kalmus, Veronika; Talves, Kairi . "Girls' and boys' experiences of online risk and safety". In: Carter, Cynthia; Steiner, Linda; McLaughlin, Lisa (Ed.). The Routledge Companion to Media and Gender (190−200), Routledge, 2014, p 192.

[106] OFSTED, "The Safe Use of New Technologies," 2014, https://webarchive.nationalarchives.gov.uk/20141105221831/https://www.ofsted.gov.uk/sites/default/files/documents/surveys-and-good-practice/t/The%20safe%20use%20of%20new%20technologies.pdf, p 4 (last accessed 9 July, 2020).

[107] Google Public Policy Blog, "Continued progress on fighting piracy", 17 October 2014, https://publicpolicy.googleblog.com/2014/10/continued-progress-on-fighting-piracy.html (last accessed 9 July 2020).

Such restrictions impact freedom of expression, the right to privacy, non-discrimination, freedom of assembly and the right to effective remedy, while failure to adequately address abusive or illegal behaviour may equally undermine a range of human rights.[108]

1. Freedom of expression

Due to the involvement of a state, a co-regulatory system that restricts freedom of expression must respect minimum criteria, as laid down in Article 10.2 of the European Convention on Human Rights.

Questions to be considered when assessing the legality of any such restrictions include the following:

- Was there an interference with the right in question and, if so, was it prescribed by law?
- Was it genuinely in pursuit of one or more of the legitimate aims at issue?
- Taking all the relevant circumstances into account, was it necessary and proportionate in a democratic society for these ends?[109]

Due attention should also be given to the fact that the restrictions imposed by content moderation are taking place in an environment that is already challenging for the exercise of this right, due to the chilling effects of pervasive online profiling.

While, as mentioned, self- and co-regulatory approaches have the advantage that they can be designed and implemented quickly, this speed comes at the expense of the deliberation and checks and balances of a legislative or judicial procedure. Such speed should not come at the cost of reviewing fundamental questions such as these. Where speed is essential, clear, predictable and accountable rules should be in place to ensure that restrictions can be temporarily enforced pending a final assessment.

The European Commission's ostensibly self-regulatory Code of Conduct on Countering Hate Speech Online addresses this issue explicitly, stressing the need to defend freedom of expression, using wording from case law of the European Court of Human Rights. Unfortunately, however, no visible effort has been made in implementation reports to assess the extent to which this provision was, in fact, respected.

Nonetheless, it is very positive that the European Commission took the initiative to prepare "evaluation reports" to monitor the effect of the Code, even though the metrics used are quite limited.[110]

It is essential, both before and after a co-regulatory measure is implemented, that the compliance with human rights law and principles be reviewed.

Case study: Removed pages / shadow bans

In the UK, the public Facebook forums ("pages") of eight independent civil society organisations were removed by Facebook on 4 November 2019 (during a general election campaign). The

[108] As just one of many examples, Netflix allegedly used copyright takedown requests to silence criticism of a controversial film, for example. See Katie Cox, "Netflix files copyright claims against tweets criticising movie, trailer", Arstechnica.com, 11 May, 2020, https://arstechnica.com/tech-policy/2020/11/netflix-dmca-takedown-requests-hit-negative-tweets-about-cuties/ (last accessed 18 January 2021)

[109] Steven Greer, "The Exceptions to Articles 8 to 11 of the European Convention on Human Rights," Council of Europe Publishing, 1997. https://www.echr.coe.int/LibraryDocs/DG2/HRFILES/DG2-EN-HRFILES-15(1997).pdf, last accessed 28 August 2020.

[110] The monitoring reports are available at https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en (last accessed 05 January 2021)

common features of all of the groups are that they started as pro-EU organisations, are all local, volunteer-based groups, based in individual towns or cities and that their local, pro-EU focus is clear from their names ("Banbury for Europe," for example). The groups also generally engaged in offline campaigning.

Some of these groups were also the subject of repeated "shadow bans" (which leave the content/accounts of the groups online but render them significantly more difficult to find) in November and December of that year. The impact of these measures was a reduction in "daily reach" of the pages of over 90%. Facebook's actions had an unknowable impact on the actions and success of the groups in relation to their influence on the election in the constituencies in which they were active.

No accusation of illegal activity was made, nor did Facebook make any specific allegation of breaches of its terms of service. Facebook suggested that the groups alter their behaviour, such as by reducing the number of posts on the pages. However, Facebook did not say if this would, in fact, stop the same problem recurring. The company explained that this restriction on the groups' freedom of expression "are taken automatically by our [artificial intelligence] AI as a result of activity undertaken by the page".

In the context of an online platform that is now so pivotal for online communications, this raises multiple questions regarding the positive obligations on the state to ensure freedom of expression, and the minimum levels of transparency, foreseeability, fairness and redress that can be reasonably expected from an internet platform.[111] It also raises another issue – if we expect artificial intelligence to be good enough to interpret complex language and its context and make decisions based on this analysis, we must demand that it be good enough to automatically provide meaningful data on the basis for this interpretation.

2. Right to privacy

Content moderation requires the processing of a range of personal data. For example, in order to implement measures such as YouTube's "three strikes" policy, a range of personal and non-personal data must be stored by the company, such as the username of the individual, the name of the complainant, the justification for the removal of the content, dates and times of uploads and removals and so on.

Furthermore, the processing of such data may include the processing of special categories of data such as in relation to political opinions, trade union membership, religious or other beliefs. Such data may only be processed under Convention 108+[112] if appropriate safeguards exist in law. It would be valuable for Council of Europe member States to review if there are specific legal grounds for processing personal data in relation to all aspects of content moderation.

[111] Monica Horten, "Algorithms Patrolling Content: Where's the Harm?", February 2020. https://ssrn.com/abstract=3792097

[112] Amending protocol to the Convention for the Protection of Individuals with Regard to the Processing of Personal Data, adopted by the Committee of Ministers at its 128th Session in Elsinore on 18 May 2018, https://rm.coe.int/convention-108-convention-for-the-protection-of-individuals-with-regar/16808b36f1

The right to privacy of complainants needs to be given particular attention, particularly in relation to reports from vulnerable individuals or groups. Failure to do this can result in a backlash directed at those individuals and groups causing direct harm to them and a chilling effect on future reporting.

Separately, while providing as much transparency to users as practical, particular attention would need to be given to any placeholder text/images that appear in place of content that has been removed. In the absence of a legally binding decision, it may not be appropriate to use wording that could be understood as an accusation of law-breaking, particularly if the law being applied is the law of a country other than the place of residence of the uploader. In Europe, Google tags all searches for personal names (but only when both the first name and surname are used) with a confusing and unnecessary notice saying "some results may have been removed under data protection law in Europe," which provides no useful information to those reading it, such as why this might be the case or what the likelihood is this is the case. Any notifications need to be clear and meaningful for users.

Case study: Placeholders and data protection

As an example of the problems that need to be avoided, the experience of users of the web hosting service mooo.com is worth noting. Mooo.com was a website hosting service, where each hosted site was a sub-domain of the service's domain name (hostedsite.mooo.com, for example). A small number of the hosted sites was discovered to contain illegal material.[113] Instead of seizing the small number of sites containing illegal material, the US authorities seized the entire mooo.com domain and replaced everything it hosted with the image below. As a result, visitors to any of the 84,000 legal websites hosted on the service saw the image, even though the site they were visiting had never contained illegal material.

Figure 2 Image displayed to visitors to tens of thousands of entirely legitimate websites hosted on mooo.com, as a result of the domain being seized by US law enforcement authorities.

[113] Ernesto Van Der Sar, "US Government shuts down 84,000 websites 'by mistake", torrentfreak.com, February 16 2011, https://torrentfreak.com/u-s-government-shuts-down-84000-websites-by-mistake-110216/ (last accessed 27 May 2020).

3. Freedom of assembly and association

Issues around content moderation and freedom of assembly are explained in General Comment 37 of the United Nations Human Rights Committee.[114] It explains in paragraph 9 that "full protection of the right of peaceful assembly is possible only when other, often overlapping, rights are also protected, notably freedom of expression, freedom of association and political participation." It importantly also points, in paragraph 34, to the positive obligation of States to ensure that "internet service providers and intermediaries do not unduly restrict assemblies or the privacy of assembly participants," which would cover both self- and co-regulatory measures.

Freedom of assembly and association can be undermined in two ways by inappropriate content moderation, namely offline or online. The online planning of physical demonstrations can be obstructed and, secondly, assembly and association online can be restricted.

Case study – freedom of assembly and association:

Four major climate organisations (and, allegedly, hundreds of others) were blocked from sending/receiving messages on Facebook the day before a planned online action against a specific investment firm. The groups had previously protested online against the same firm.

Facebook initially claimed that the organisations had committed an intellectual property infringement before, then claiming it was a mistake and restoring the accounts gradually, after the date of the planned protest. [115]

In order to fulfil their positive obligations to ensure freedom of assembly and association and, indeed, for freedom of expression more generally, it is necessary that states equip themselves with the requisite legal powers and that self- and co-regulatory agreements are designed in a way that avoids the implementation of unjustified restrictions. Appropriate and dissuasive redress rules and penalties should be in place to discourage internet intermediaries from making mistakes.

Freedom of assembly and association are already threatened by an environment where moderation and curation of content is carried out by often opaque artificial intelligence systems, meaning that the "public space" may be different for everyone. Common points of reference are rendered less easy to identify, with opinion forming being partly influenced by this technology.

If content moderation deliberately or accidentally reduces the diversity of information available to individuals or groups, this also undermines their freedom of assembly and association. It is further important to note that the largest internet intermediaries now sell micro-targeted election influencing as a service. This raises significant questions for democracy and possible conflicts of interest.[116] This

[114] United Nations Human Rights Committee, "General comment No. 37 (2020) on the right of peaceful assembly (article 21)," September 2020. https://tbinternet.ohchr.org/_layouts/15/treatybodyexternal/Download.aspx?symbolno=CCPR%2fC%2fGC%2f37&Lang=en (last accessed 25 September, 2020)

[115] Oliver Milman, "Facebook suspends environmental groups despite vow to fight disinformation" The Guardian, September 2002, https://www.theguardian.com/environment/2020/sep/22/facebook-climate-change-environment-groups-suspended (last accessed 8 October 2020).

[116] Zuiderveen Borgesius et al, "Online Political Microtargeting : Promises and Threats for Democracy," Utrecht Law Review, Volume 14, Issue 1, 2018 https://www.ivir.nl/publicaties/download/UtrechtLawReview.pdf (last accessed 28 August 2020)

might lead to situations where intermediaries may be minded to be more lenient in relation to political speech which comes from a politician who pays for such targeting or who generates controversy, engagement and, therefore, revenue.

Case Study: Law, politics and self-regulation in the Irish abortion referendum

Ireland held a referendum on the legalisation of abortion on 25 May 2018. Ireland has extensive rules on spending in election campaigns, with different rules on referendum spending. Rules on referendum campaign funding focus on groups that receive funding, while nominally "self-funded" groups are not subject to the same level of scrutiny.

There is no direct regulation of online advertising even though there are rules on campaign posters and a ban on paid TV and radio advertising. The government of Ireland had not updated the law to change this (although successive governments since 2008 have planned to establish an electoral commission to address these problems and this process is moving forward apace at time of writing), leaving a disparity between offline and online advertising rules.

This unclear situation between the treatment of funded and self-funded groups and between offline and online rules ultimately led to unilateral action being taken by internet intermediaries, who are private corporations that were receiving money from campaign groups. This action had an unknowable impact on the actions and success of the groups. In retrospect, this could have been avoided if the State had taken action in advance to ensure that the rules were clear and equitable.

In the months leading up to the referendum, Google and Facebook accepted paid-for advertising for both the "yes" and the "no" campaigns. Twitter, on the other hand, refused all advertising, remaining consistent with their rules on advertising related to medical services and their rules on political advertising.

On Tuesday, 7 May 2018, Facebook took the self-regulatory decision to ban all advertising in relation to the referendum from advertisers based outside Ireland.

On Wednesday, 8 May 2018, Google took the self-regulatory decision to ban all advertising in relation to the referendum, regardless of who paid for it.

This example illustrates the need for accountability in content curation and, as appropriate, state guidance. It also illustrates that the relevant decisions taken by internet intermediaries can have financial consequences for them, both positive and negative.

4. Right to remedy

In general, the emphasis should be on avoiding the necessity to resort to redress mechanisms for incorrect decisions in self- or co-regulatory content moderation. This means ensuring that all reasonable measures are taken by each stakeholder to ensure that rights are not violated. When remedies are needed, all necessary tools must be made available, such as legal aid and victim support and access to due process.

The balance of incentives for providers needs to be such that redress, and information as to how it can be accessed, is made readily available.

a) For victims

Apart from bringing an end to the infringement or offence, there is little that a self- or co-regulatory approach to content moderation can do to provide redress for victims of illegal online behaviour or

offline crimes.[117] It is therefore crucial that any approach to content moderation foresee adequate cooperation with, and action by, state authorities, to ensure redress for victims – which means providing full support to negate or mitigate the damage that has been caused. The positive obligations of the State in this context were made clear by the European Court of Human Rights in the case of K.U. vs Finland, when rejecting the notion that redress could have been sought from the internet intermediary through whose services the offence in question had been committed as an alternative to pursuing the actual infringer:

"*It is plain that both the public interest and the protection of the interests of victims of crimes committed against their physical or psychological well-being require the availability of a remedy enabling the actual offender to be identified and brought to justice, in the instant case the person who placed the advertisement in the applicant's name, and the victim to obtain financial reparation from him."*[118]

b) For people whose content has been unjustly removed

It is essential that any self- or co-regulatory projects have adequate safeguards to ensure that content is not suspended or removed unjustly. In keeping with the logic of rights being the rule and restrictions being the exception, content should be left online if possible and only suspended or removed in exceptional cases where this is necessary. What is possible depends greatly on the type of content. If it is necessary to take content offline quickly, this should of course be possible, as long as the process remains predictable, necessary and proportionate.

However, it is inevitable that mistakes happen. Ensuring that there is adequate transparency to prevent similar mistakes in the future must be considered a fundamental part of any redress scheme. In the same vein, adequate transparency is essential to identify possible discrimination in the way complaints about different types of content are handled. Redress could be decided by the courts but also, if all parties agree, by alternative dispute resolution mechanisms or e-courts.

When content is removed, it is also important that transparency measures make clear the specific reasons why the content was removed as, without this, it is difficult for individuals to know if an appeal is worthwhile or possible., The right to a timely, accessible, and fair appeals process should always be accorded. The right to effective remedy does not automatically require judicial adjudication (as users and intermediaries may agree to use alternative dispute resolution mechanisms), but this should always be an option.[119]

It is simple, but misleading, to assume that the costs of content moderation mistakes are always readily quantifiable in financial terms. Human rights are priceless, as are democratic values. It is helpful to reflect on the harm caused by unjustified restrictions on human rights in this way rather than through the prism of a potential financial loss when considering redress. As the harm caused to the people whose content is unjustly removed is frequently not a clearly financial one, relying on simple calculations of possible directly attributable losses caused by the unjustified removals will generally not compensate the individuals concerned. This will also not compensate society for the chilling effect

[117] That said, attention should be given to making sure that content moderation is as sensitive as possible to the needs of those making reports in such contexts. This is discussed extensively in Alexander Brown, "Models of Governance of Online Hate Speech", independent research carried out for the Council of Europe, https://rm.coe.int/models-of-governance-of-online-hate-speech/16809e671d (last accessed 18 January 2020), pp 155-167.

[118] K.U. vs Finland, application number 2872/02, Paragraph 47, https://hudoc.echr.coe.int/eng#{%22appno%22:[%222872/02%22]} (last accessed 04 June, 2020).

[119] See, in particular, Paragraph 1.5.2 of Recommendation CM/Rec(2018)2 of the Committee of Ministers of the Council of Europe on the roles and responsibilities of Internet intermediaries, https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680790e14 (last accessed 21 May 2020).

of such restrictions. Furthermore, such normally minimal financial costs will not serve to incentivise intermediaries to take more care in the future. At the same time, intermediaries should, in turn, have the right to meaningful and dissuasive redress against negligent or wilfully incorrect reports of prohibited content from individuals or organisations. Penalties for damage caused by excessive content moderation must therefore take the broader harms to society into account and must be dissuasive rather than just compensatory.

The role of the State in ensuring that business-related human rights restrictions are remedied was clearly highlighted in the UN Guiding Principles on Business and Human Rights:

> *"Unless States take appropriate steps to investigate, punish and redress business-related human rights abuses when they do occur, the State duty to protect can be rendered weak or even meaningless."*[120]

c) Adequacy of the right to redress and remedy

In cases where the restricted content or services have caused a demonstrable negative impact to a significant section of the public, for example the spread of dangerous disinformation,[121] mechanisms should be in place whereby this is made good by equivalent messages being communicated to equivalent numbers of people or, if the data is available, the same people, in the same format and frequency.[122] Publishing a correction that will be seen by ten people as redress for a message seen by ten million people is not an adequate response.

IV Purposes and drivers of content moderation

This chapter explores the purposes and drivers of content moderation. This is because, when developing policy regarding content moderation, it is crucial to take into account the fact that the reasons for content moderation, as well as the types of content that are moderated, vary considerably. In order to develop effective policy, it is essential to understand the type of content being regulated and the intended outcomes.

This chapter looks at content moderation from two perspectives, namely when it is implemented for business reasons and for public policy reasons.

1. Content moderation & business interests

In order to adequately protect human rights in relation to content moderation, it is important to understand that, seen from the perspective of the intermediaries, this activity has multiple drivers. It can be undertaken:

- to ensure that the business model remains valid – a car-focussed platform may wish to identify and remove content associated with subjects other than cars;
- to ensure that there is no content on the platform which would be displeasing for advertisers and therefore reduce revenue;
- to avoid liability for potentially illegal content or

[120] United Nations, "UN Guiding Principles on Business and Human Rights," 2011, https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf (last accessed 2 September 2020)
[121] The ongoing work of European Regulators Group for Audiovisual Media Services is significant in this context. See https://erga-online.eu/ for more information. (last accessed 12 October 2020)
[122] To be effective, a remedy must be capable of directly providing redress for the impugned situation (Pine Valley Developments Ltd and Others v. Ireland, Commission decision, 1989).

- to be seen to be making an effort for the betterment of society by fighting potentially illegal content.

Problematic content exacerbated by business models

Many internet intermediaries have a business model that is partly or fully reliant on the collection and use of personal data from users. This data is generated when users are engaged with content on the platform. The design of the intermediary's service can serve to drive such engagement (by rewarding people who post content that drives engagement with "likes" or similar positive feedback).

As a result, internet intermediaries are faced with a potential conflict of interest if content that leads to higher levels of engagement is also content which is illegal or otherwise unwelcome. One, albeit small-scale, study found that, at one moment, a quarter of the top viewed videos[123] on YouTube on Covid-19 contained misleading information.[124] Potential conflicts of interest must be identified and mitigated when planning any self- or co-regulatory scheme. If illegal or unwelcome behaviour is being driven by the business model and interests of a platform, then this is the problem that needs to be solved, and not how quickly the behaviour can be deleted or demoted. While exceptions may exist, a self- or co-regulatory scheme should not rely on online intermediaries acting in a way inconsistent with their own financial interest or the foundations of their business model.

In some cases, social media companies are also involved directly in online tracking on, for example, news sites. So, the spreading of a sensationalist or misleading news story may bring the internet intermediary revenue both from the data this story generates from the original news site and from the engagement the story generates when it is posted on their platform, creating a second layer of conflict of interest.

The scope for potential conflicts of interest is all the greater when additional economic interests come into play. Many online internet intermediaries also sell economic and political influence by microtargeting consumers and voters.

By contrast, decentralised social media platforms, such as Mastodon, allow small and large communities to set and manage their own moderation policies, and give users more choice over whom they follow across these communities, than centralised platforms.[125] Importantly, decentralisation helps minimise one of the major problems in the current "market", namely content moderation at scale. On the one hand, Mastodon, being decentralised and open source, allows heavily regulated safe spaces for interaction but, on the other, cannot prevent unregulated spaces being set up. However, this happens without the risk of amplification to other audiences, as tends to happen in centralised, data-driven platforms. The German Federal Data Commission launched its own instance of Mastodon in 2020.[126]

In conclusion, the changing, and possibly contradictory, incentives of market players, and the consequences of the design of online services, must be fully understood and taken into account when self- and co-regulatory projects are being designed or called for by public authorities, in order to ensure predictability, proportionality, legitimacy and effectiveness.

[123] English language, non-duplicate videos of less than one hour in length.
[124] Li HO, Bailey A, Huynh D, et al, "YouTube as a source of information on COVID-19: a pandemic of misinformation?" BMJ Global Health 2020;5:e002604, https://gh.bmj.com/content/5/5/e002604 (last accessed 20 May 2020).
[125] For a more in-depth analysis of issues around decentralisation and interoperability, see: Ian Brown, "Interoperability as a tool for competition regulation" 30 July, 2020. https://osf.io/preprints/lawarxiv/fbvxd/ (last accessed 09 October 2020)
[126] See https://social.bund.de/@bfdi/105026921216079123 (last accessed 13 October 2020)

The case study above on the Irish abortion referendum illustrates that there are sometimes no neutral decisions, with even complete inaction by the state having a significant impact on the democratic process.

2. Content moderation for public policy reasons

In order to develop effective policies to address problematic content online, it is crucial to understand the nature of such content. It is unlikely to be the case that radically different problems will require identical solutions. The following categorisations are intended to illustrate the range of content that is being addressed and is not meant to be definitive or immutable. Some types of content may fall into multiple categories. States should ensure that the role that intermediaries are expected to take in the fight against illegal online content is adapted to the type of illegal content in question.

a) Content that is illegal everywhere, regardless of context

There are very few examples of content that is illegal everywhere. The clearest example is child sexual abuse material.[127] Depictions of child abuse (sometimes referred to as "child pornography in older texts," [128]) are prohibited in the Council of Europe region and beyond by Article 9 of the Budapest Convention, Article 20 of the Lanzarote Convention and by several international legal instruments, such as Convention 182 of the International Labour Organisation, the UN Convention on the Rights of the Child, and others.

The approach to this problem in Europe (notice and takedown, using hotlines, with some countries implementing various types of web blocking) has been almost static in the past fifteen to twenty years, with no meaningful assessment of the effectiveness of the measures in force nor of how the crime has evolved. Issues surrounding if, how, for how long and by whom, with what levels of transparency, data should be stored in this context are only now beginning to be discussed. For such crimes, continuous assessment is indispensable in order to ensure ongoing effectiveness of measures taken to fight them.

b) Illegal content that is part of a wider crime

Such offences are generally more serious and more urgent. For such material to be available online, such as the offer for sale of products made from a protected animal species, at least one additional crime is likely to have taken place. Any content moderation initiative that fails to take the offline elements of a crime into account risks leaving victims without redress. On its own, there is nothing that a self- or co-regulatory scheme can do to investigate or dissuasively sanction the offline elements of these crimes.

It is therefore crucial that any self- or co-regulatory initiative to fight serious crime (that constitutes a threat to human life or child abuse, for example) include responsibilities for states to take all necessary measures to address the wider problem. It should never be possible to adopt a self- or co-regulatory approach in relation to such content without explicit reference to the expected engagement with law enforcement and other relevant state authorities. For example, the German "Network Enforcement Law" ("NetzDG") requires some associated data to be stored in case this is needed for subsequent investigations. Such measures need to be calibrated very carefully in order to avoid unintended consequences for privacy and other human rights.

[127] Even here, laws are not always uniform, with flexibilities available to states with regard to, for example, national legislation on age, apparent age, possession for private use, creation and possession by children under instruments such as the Lanzarote Convention (Treaty 201 of the Council of Europe) and the EU Directive 2011/93 on combating the sexual abuse and sexual exploitation of children and child pornography.

[128] This term has fallen out of favour and is now generally avoided.

c) Content that is not necessarily part of a wider offence,

At the other end of the spectrum, some online offences may have no offline component. For example, if somebody uploads a copy of a film to their website or to a social media service, the content itself is not illegal, the act of uploading the content may or may not be illegal, may be subject to legitimate exceptions and limitations of copyright and, indeed, if nobody actually downloads the film, no actual prejudice was suffered by the owner(s) of the rights to the film. As a result, content moderation that focuses on availability will address that issue in a more comprehensive way than it would if it were an offence with an offline component. If availability is the only problem, removing availability solves it (albeit with the risk of creating further problems for freedom of expression, right to redress, etc). Conversely, if availability is not the only problem, it will not be solved by removing availability.

d) Legal content that is illegal primarily due to its context

This would include, for example, so-called "revenge pornography," or unauthorised publication of personal information, where the content itself is not illegal, but the manner in which it becomes available is an offence. [129] This can be difficult, if not impossible, to identify based solely on an appraisal of the content in question. On the other hand, "doxing" on social media for example, can sometimes be very easy to identify.

e) Content that is illegal primarily due to its intent.

This would include, for example, incitement to violence or incitement to terrorism. It is not the words themselves, but rather the intent, content and status of the speaker that leads to the offence being committed. [130] [131] [132]

f) Content that is potentially harmful but not necessarily illegal

This is a very wide category. It covers, for example, legal but possibly dangerous support or advice related to self-harm or suicide. Depending on their business model or intended clientele, internet intermediaries may choose to try to restrict such content. Such restrictions may not breach freedom of expression as long as information or ideas "that offend, shock or disturb the State or any sector of the population" have adequate avenues to be expressed. This principle was stressed by the IT companies and by the European Commission on the first page of their "Code of Conduct on Countering Illegal Hate Speech Online" and is a recognition by the participating companies of their role in upholding key human rights values in the context of co-regulatory schemes.[133]

The task, as always, should be to define the most desirable outcome, which involves identifying the type of content, the audience for which it is considered harmful and the nature of the feared harm.

---

[129] The Rabat Plan of Action establishes six criteria on "Freedom of Expression vs incitement to hatred" to establish a "high threshold for restrictions on freedom of expression". These are 1. The social and political context, 2. The speaker's position or status, 3. The intent to incite audience against target group, 4. The Content and Form of the Statement, 5. The Extent of the Dissemination, 6. The Likelihood of Harm, including imminence. https://www.ohchr.org/EN/Issues/FreedomOpinion/Articles19-20/Pages/Index.aspx (last accessed 18 January 2020). Similar criteria have been used by the European Court of Human Rights, for example in Leroy v. France (application no. 36109/03), Jersild v. Denmark (application no. 15890/89, Feret v. Belgium (application no. 15615/07) and Vejdeland and Others v. Sweden (application number 1813/07.

[130] Ibid

[131] See also, Article 19, "Hate Speech Explained: A Toolkit", 2015. https://www.article19.org/data/files/medialibrary/38231/'Hate-Speech'-Explained---A-Toolkit-%282015-Edition%29.pdf (last accessed 04 January 2021)

[132] UN Committee on Elimination of Racial Discrimination, "General Recommendation 35 Combating Racist Hate Speech," 2013, p.5, https://www.refworld.org/docid/53f457db4.html (last accessed 19 January 2021).

[133] Code of Conduct on countering illegal hate speech online, 30 June 2016, https://ec.europa.eu/newsroom/just/document.cfm?doc_id=42985 (the link opens a Word document) (last accessed 21 May 2020).

g) Content that raises political concerns

Any of the above categories could, at a given moment, fall into this category. It is the nature of media that some types of content suddenly become the topic of media attention, without necessarily being, of itself, a problem that needs to be, or can be effectively, addressed by content moderation or regulation. For example, "deep fakes" may not reach a level that requires a policy intervention, but one high-profile case may create an impetus to act. Politically or for company PR needs, it is easy to get caught in a spiral of knee-jerk reactions because "someone should do something". Having a clear, predictable approach to building proportionate and effective responses should mitigate the pressure for such knee-jerk reactions, while improving the quality of those responses.

3 Conclusion

Clearly, these types of content are fundamentally different, not just in terms of their illegality, but also their characteristics and the gravity of their consequences. Treating a copyright violation in the same way as an upload of child sexual abuse material, is unlikely to be appropriate for at least one, if not both, types of content. It is therefore crucial to tailor content moderation responses to the specific problem they are trying to solve.

It should also be noted, of course, that laws can change, so content that is legal in one country on one day can become illegal the next day and vice versa.

V. Structures for content moderation

Self- and co-regulation are common and often highly successful approaches to addressing policy concerns across wide swathes of industry. However, these terms are understood in different ways in those different contexts and have different impacts in different environments. This section looks at the concepts of self- and co-regulation before drawing some conclusions from experience regarding common characteristics that can be found in some successful approaches.

1. Self-regulation

Self-regulation is the regulation of a company or a sector of itself in order to achieve an industry or public policy objective. It can also be implemented as a strategy to avoid traditional regulation.

Wide-ranging successes of self-regulation in the media sector have shown the positive impact that this approach can have in certain circumstances. When self-regulation has worked well, it has led to the creation of, for example, ethics codes, ombudspersons and innovative complaints mechanisms that permit news media to remain independent while maintaining high standards. Generally, there is a strong incentive (such as upholding high editorial standards, reputation, flexibility, and independence) for media companies to ensure the success of such initiatives. Of course, where those incentives do not exist, due to state interference or lack of independence from political influences, such incentives are less present.

In principle, some of the same advantages can apply in relation to moderation of content online, with flexibility being a key advantage. If an internet intermediary is, in fact, regulating itself (such as choosing to permit political advertising or not), and if it has a clear business interest in ensuring effectiveness (such as filtering unsolicited advertising via e-mail) and if the endeavour has a perceived benefit for its users, it can work well. The converse is also true, when the incentives of the provider are not aligned with the interests of the regulator or users, self-regulation is less likely to work well.

Sometimes internet intermediaries may have conflicting interests, for example their interest in protecting their users from misinformation, but also an interest in gaining revenue from the controversy and engagement that may come from the spreading of such types of content.

Furthermore, it is common for internet intermediaries to have multiple business models running in parallel. A social media provider may make no revenue at all from its social media function, but from advertising, collecting user data or from merging that data with data from third parties. This means that the balance of incentives for the company in ensuring the success of any such initiative may be evolving and unpredictable, even to the company itself.

The term "self-regulation" should only be used to refer to situations where a company or group of companies are acting to regulate their own activities, without direct or indirect state pressure. It is important to be clear about terminology, as the degree of state involvement is significant for the legal responsibilities of the State.

An internet intermediary can be involved in a wide range of types of self-regulation from small scale (whether or not to accept political advertising), to large scale (policing all potentially harmful expressions from its users). Such self-regulation can range from self-interested (preventing spam) to possibly not self-interested (preventing content that could be profitable). This means that few, if any, assumptions should be made about the desirability of self-regulation in relation to content moderation by internet intermediaries.

2. Co-regulation

Where the state and private actors cooperate to create an ad hoc framework to address a public policy problem, this is referred to as "co-regulation". It can also cover situations where industry associations adopt codes, adherence to which can be used to demonstrate compliance with legal obligations.

The European Court of Human Rights has stated that self- and co-regulatory mechanisms may be acceptable, on the condition that they include effective guarantees of rights and effective remedies for violations of rights.[134] For co-regulatory measures, the Court demands a considerable degree of government involvement such as the approval of the rules.

Co-regulatory mechanisms should be based on a legal framework set up by the State. Such a framework should define clear limits and provide safeguards to prevent arbitrary decisions by non-state agents.[135]

Generally, the same considerations apply to the enforcement of co-regulatory approaches as to self-regulatory approaches. However, co-regulatory approaches allow a diligent public authority to require more transparency and more accountability than might otherwise be the case.

For example (and even though this co-regulation is referred to by the European Commission as self-regulation), the "Guiding Principles" on "The Follow the Money Approach to IPR Enforcement" are a good example of the kinds of safeguards that can be envisaged."[136] This agreement was negotiated between the European Commission and industry and civil society stakeholders. It promised that the ensuing Memorandum of Understanding would develop key performance indicators, an independently assessed "verification and compliance" process and that an appropriate balance between the various fundamental rights at stake would be ensured and demonstrated. Such an

[134] ECtHR, Peck v. the United Kingdom, no. 44647/98, 2003-I, paras 108 and 109

[135] Aleksandra Kuczerawy, "Private enforcement of public policy: freedom of expression in the era of online gatekeeping," PhD thesis, KU Leuven, 2016.

[136] https://ec.europa.eu/docsroom/documents/19462 (last accessed 12 May 2020).

approach would allow for protection for human rights, accountability, flexibility, and meaningful transparency, as well as independent verification that the public policy objectives were also being achieved.[137]

It should be noted, however, that in reality self- and co-regulation generally do not, in fact, exist as two clear, discrete concepts. Such initiatives often come into existence due to government pressure, a perception on the part of industry that legislation is pending and should be pre-empted, or overt "threats" of legislation by governments. In relation to the "follow the money" projects of the European Commission, it announced in 2015 that legislation may follow if the "self-regulatory" approach was not fully effective. Furthermore, despite calling for, convening and negotiating the "follow the money" code of conduct for advertisers and despite the "guiding principles" of the code giving the Commission specific tasks, the European Commission still refers to the instrument as self-regulatory".[138]

When public policy priorities are at stake, there seems little value in adopting an approach which is either purely self-regulatory or an unclear mix of self- and co-regulatory. A clear, target-driven and accountable approach based on constructive engagement by state authorities is more likely to achieve positive results.

3. Common characteristics of successful approaches

Much can be gained by reviewing the history of self- and co-regulation, including in other industries, to help avoid pitfalls and to adopt practices that maximise chances for success, while remaining cognisant of the particular importance of online communications tools for democratic processes and institutions, peace and stability, equality and non-discrimination.

One study compared the characteristics and relative successes and failures of self-regulatory regimes in the forestry, fishing, tobacco, soft drinks and fast-food industries.[139] Looking at the relative successes of various schemes, the researchers determined that the motivation behind the initiatives may be key:

> *"The type of motivation may be a determinant of success. In some cases, an industry perceives that it must police itself because governments are involved too little, as was the case with forest and fisheries stewardship. For other industries, government intervention is perceived as a threat, and self-regulatory actions are a means to prevent or forestall outside regulation."*

The former motivation, a clear business need, in the absence of government leadership, appears to be a determinant of success across all the schemes examined. However, self-regulation as a means of forestalling government intervention was a determinant of failure. This appears logical, as the direct purpose of the self-regulatory initiative is to forestall regulation and not to solve the public policy problem itself.

[137] Four years after the adoption of the "guiding principles", the European Commission is unable to say if these commitments were actually respected. See parliamentary question 3115/2020 by Patrick Breyer MEP to the European Commission, https://www.europarl.europa.eu/doceo/document/E-9-2020-003115_EN.html (last accessed 1 September 2020)

[138] See https://ec.europa.eu/growth/industry/policy/intellectual-property/enforcement/memorandum-of-understanding-online-advertising-ipr_en for example (last accessed 12 May, 2020.

[139] Sharma, L. L., Teret, S. P., & Brownell, K. D. (2010). "The food industry and self-regulation: standards to promote success and to avoid public health failures." American journal of public health, 100(2), 240–246, 2011. https://doi.org/10.2105/AJPH.2009.160960 (last accessed 8 October 2020)

Based on the experience of the various self-regulatory initiatives examined, the authors list nine key standards on which future self-regulation in the food sector could be built, learning from the mistakes and missteps of the past.

Aim	Standard
Transparency	1) Transparent self-regulatory standards created by a combination of scientists (not paid by industry) and representatives of leading nongovernmental organizations, parties involved in global governance (e.g., World Health Organization, United Nations Food and Agriculture Organization), and industry
	2) No one party given disproportionate power or voting authority
Meaningful objectives and benchmarks	3) Specific codes of acceptable behaviours based on scientifically justified criteria
	4) Predefined benchmarks to ensure the success of self-regulation
Accountability and objective evaluation	5) Mandatory public reporting of adherence to codes, including progress toward achievement of full compliance with pledges and attainment of key benchmarks
	6) Built-in and transparent procedures for outside parties to register objections to self-regulatory standards or their enforcement
	7) Objective evaluation of self-regulatory benchmarks by credible outside groups not funded by industry to assess health, economic, and social outcomes
	8) Periodic assessments/audits to determine compliance and outcomes
Oversight	9) Possible oversight by an appropriate global regulatory or health body (e.g., World Health Organization)

A similar analysis of self-regulatory initiatives in relation to content moderation and self- and co-regulation in the internet sector appears to be very overdue and likely to generate similar results. This assessment broadly overlaps with that of the UK's National Consumer Council. The latter found that self-regulatory initiatives must have clear policy objectives, must not inhibit the scope for competition, must have a strong independent element for both design and governance, must have a dedicated institutional structure and that they should operate within a clear legal framework.[140]

VI. Transparency

Why transparency is essential

Transparency has two components, the first relates to what is being done or attempted (namely seeking to impose clear terms of service or a specific law). The second, which is the subject of this section, relates to the effects of what is being done – details on what was removed, how much was removed based on what grounds, how much was put back, the number of complaints received regarding infringing content or takedowns, the impact on the problem(s) being addressed and so on. Where possible, such data should be produced using standardised methodology and in machine-readable formats.

[140] Chris Jay Hoofnagle, "Federal Trade Commission Privacy Law and Policy" Cambridge University Press, 2016. Chapter 6 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2800276 (last accessed 28 May 2020).

Particularly in a highly fluid environment where both crimes and technologies change continuously, and where the border between State and private actions is often unclear, it is impossible to ensure predictability, necessity and proportionality on an ongoing basis without the data to carry out such assessments.

The cost of self-regulation and co-regulation is often a reduction in accountability and democratic legitimacy. The key benefit of self- and co-regulatory approaches in general is that they are flexible. This benefit is particularly valuable in the ever-changing online environment. As problems change, responses, logically, also need to change. However, without meaningful transparency, it is not possible to identify and assess changes nor make corresponding adjustments. Therefore, without transparency, society loses one of the key benefits of self- and co-regulation, while still incurring their cost.

To ensure restrictions are necessary and proportionate

Restrictions on human rights must be necessary and proportionate when they are launched, but also on an ongoing basis. In Belgium, a code of practice was signed whereby individual "newsgroups"[141] were removed from a participating internet intermediary's servers, if a particular interest group requested their deletion due to alleged illegal sharing of copyrighted material.[142] In response, users monitored which newsgroups were available from day to day on the services of participating internet intermediaries. They did this to identify which newsgroups had been removed and, by extension, which newsgroups were "best" for finding unauthorised content. They then published the lists of the newsgroups online, enabling them and others to use alternative providers to access the content in question.[143] The project had become counterproductive in a matter of days.

It was very clear at the time that the effectiveness of the measure collapsed in the few days between its launch and when users changed behaviour in response. In most cases, such evolutions will be far less visible and less sudden. Without mechanisms to identify and remedy such developments on an ongoing basis, effectiveness, necessity, and proportionality over time cannot be guaranteed. Particularly in situations where urgent action is needed, additional care must be taken to ensure that the evolution of the problems in question is carefully tracked.

To ensure non-discrimination

Discrimination can even happen in well-designed and well-intentioned content moderation systems. It can happen due to lack of local resources, lack of insights into regional use of language or different use of words by different societal groups, "gaming" (targeted manipulation) of automated systems by bad actors, and numerous other factors. It is therefore essential that transparency systems be designed to produce the data that is necessary to identify such discrimination quickly. Systems that do not have the necessary mechanisms to produce data to identify potential discrimination should not be permitted.

To ensure accountability of stakeholders (such as States)

Transparency is also essential to ensure accountability beyond the direct stakeholders (the complainant, the internet intermediary, and the individual or group whose content is subject to the complaint). This is most obviously the case when content is part of an offline crime. If evidence of a crime has been identified, data on how quickly it was deleted by the internet intermediary gives an

[141] Global discussion boards that are replicated, hosted and made available by service providers for their users. These were popular in the early days of the internet but now have fallen out of fashion.
[142] Mick De Neeve, "Belgische Providers Schrappen Nieuwsgroepen," Tweakers.net, 16 July 2005, https://tweakers.net/nieuws/38089/belgische-providers-schrappen-nieuwsgroepen.html (last accessed 12 May 2020).
[143] See, for example https://userbase.be/forum/viewtopic.php?f=24&t=7895&start=40 (last accessed 28 May 2020).

incomplete picture. Instead, data is needed on the number of reports that were made available to, or referred from, law enforcement authorities and the number of investigations and prosecutions that were launched as a result.

As an example, the main internet intermediaries use a closed-source photoDNA technology to block uploads of known child abuse material. If individuals are uploading known child abuse content, data should be collected on how the crime evolves in response to the measure, if or how law enforcement authorities have the possibility to become aware of instances of such blocking, whether and how often associated personal data is accessed by national law enforcement authorities, if data is automatically made available to law enforcement authorities or not, and if this leads to any investigations or prosecutions, in order to measure how successful the method is. Due to the seriousness of the crime in question, independent testing should be required. Deletion or blocking, as a stand-alone strategy, essentially creates impunity for serious crime.

Identification of transparency data

Rigorous and ongoing efforts are needed to identify necessary transparency data. It is crucial that all data required to measure success or failure and to identify counter-productive or negative impacts be collected. Data collection criteria must also be reviewed to ensure that they do not create perverse incentives (like incentivising speed over accuracy, which can also result in "gaming" of the complaints mechanisms to the detriment of vulnerable groups). Relevant data that should be collected can also be identified by looking at the possible risks to human rights and at the goals and minimum targets of the content moderation itself.

In the above example, measures to fight child abuse, it should be obvious that data necessary for transparency reasons would include the specific reason for the removal of the content (terms of service or law), if data is stored in relation to potentially criminal material, if this is available to law enforcement authorities on demand, if or how often this data was requested by law enforcement authorities, how often content was not immediately removed in order to avoid interfering with law enforcement investigations, other reasons for delays in takedowns, speed of takedowns/blocks, and so on.

Recognising the positive & negative incentives created by transparency metrics

If governments put pressure on internet intermediaries to take actions that they would not otherwise have taken, then those actions will be driven by the implicit or explicit targets that have been set by government demands. If the key metrics are "how many" and "how fast", the providers are incentivised to delete as much as possible, as quickly as possible.

Transparency requirements without appropriate granularity can lead to data being produced that is impossible to use, either deliberately or accidentally. For example, YouTube's transparency report[144] under the German Network Enforcement Law includes a category called "terrorist or unconstitutional content" which mixes terrorist content with breaches of provisions of the German criminal code generally, but not totally, unrelated to terrorism, such as use of symbols of unconstitutional organisations (article 86a of the Criminal Code) and certain types of forgery (article 269 of the Criminal Code). Facebook's reporting,[145] under the same provisions of the same law, is much more granular. This is to be welcomed.

[144] See https://transparencyreport.google.com/netzdg/youtube?hl=en.

[145] https://about.fb.com/wp-content/uploads/2020/01/facebook_netzdg_January_2020_english.pdf (last accessed 13 May 2020).

The divergences in methodologies leave the German state and German people without clear data that could be compared across intermediaries and over time, despite this presumably being the main reason for having transparency obligations in the first place. As the content is being removed ostensibly on the basis of specific provisions of the same legislation, it seems counterintuitive that companies are providing different kinds and detail of data. For full and effective transparency, data should be provided with maximum levels of granularity and identical methodology, allowing an effective analysis and evaluation of the content moderation methods applied. As an example of good transparency practice that should be promoted and replicated, the Austrian hotline for child abuse material and endorsement of National Socialist ideology (stopline.at) publishes transparency reports with rigorously consistent methodology. This gives policy-makers and others the data necessary to see the scale of particular problems at any given moment and to assess the evolution of the problems over time.

Finally, transparency is also needed in the promulgation and adaptation of content rules. Changes to content rules should be easy to understand and should be justified. Also, it should be made clear why particular kinds of content that are not illegal, are prohibited on the services of internet intermediaries.

Trusted flaggers

Internet intermediaries often find it useful to use specialised organisations as a filter through which to get more reliable reports of infringing content. Indeed, the role of trusted flaggers is institutionalised in the German Network Enforcement Law. These are referred to variously as "trusted flaggers" or "priority flaggers."[146] Specific transparency rules are needed for such initiatives in order to ensure that no conflicts of interest (such as structural (working on content it is trying to eliminate) or financial (being funded by the platform) conflicts) are accidentally created and that their level of effectiveness and trustworthiness remains consistently high. Use of trusted flaggers should not be mandatory and notices from trusted flaggers should not be considered "actual knowledge" of illegality of content. Doing so would give them a quasi-judicial function and may impede the use of this approach.

Oddly, trusted flaggers only exist for the restriction of content and not for its protection or putting it back online. If third parties can be trusted in a way which leads intermediaries to prioritise their notices that content is illegal and should be removed, it seems logical that third parties could be similarly trusted by intermediaries to submit priority notices that certain content or accounts should not have been restricted.

Taking this logic a step further, European public service broadcasters argue that, as they are subject to direct and comprehensive editorial responsibility, internet intermediaries should not subject them "to any form of control or interference."[147] In other words, they should be subject to a permanent "trusted de-flagging". This raises some fundamental questions, in particular, how independent would a broadcaster need to be to qualify for this status, and who would be responsible for making or reviewing a decision to award this status to a broadcaster? Furthermore, the same logic could arguably

[146] EuroISPA, "Priority Flagging Partnerships in Practice", January, 2019, https://www.euroispa.org/wp-content/uploads/Hutty_Schubert_Sanna_Deadman-Priority-Flagging-Partnerships-in-Practice-EuroISPA-2019.pdf (last accessed 28 May 2020).

[147] European Broadcasting Union response to European Commission consultation on the Digital Services Act (question 16), https://www.ebu.ch/files/live/sites/ebu/files/Publications/Position_Papers/open/EBU_response_Digital_Services_Act_consultation%2008092020.pdf (last accessed 30 September, 2020).

be used to say that any individual not speaking anonymously is subject to the law of their country which, in turn, could lead to discrimination against those who wish or need to speak anonymously.

VII. Key principles for a human rights-based approach to content moderation

1. Transparency

As explained above, transparency is the single most important element for the achievement of successful content moderation. It is essential for ensuring accountability, flexibility, non-discrimination, effectiveness and proportionality, as well as for the identification and mitigation of conflicts of interest. All the criteria listed below rely, to a greater or lesser extent, on transparency in order to be realised.

Minimum standards should be identified to assess whether the content moderation in question is achieving its specific goals. These can include standards for false negatives, false positives and response times. This means, for example, that minimum standards should be set for the amount of times that infringing content is incorrectly labelled as non-infringing and that non-infringing content is incorrectly labelled as infringing, with clearly defined standards for acceptable error rates. This requires independent review of at least a representative sample of cases. Any breach of acceptable error rates should automatically prompt corrective measures.

2. Human rights by default

Under the Convention, human rights are the default and restrictions may be exceptionally imposed when necessary and proportionate to do so. This approach must guide the development of policies in relation to content moderation.

It is also important to proactively identify which rights might be threatened before any content moderation process is initiated, while bearing in mind that failure to moderate content can undermine equality. Content moderation can, for example, restrict the rights protected by Articles 8 and 10 of the Convention. These are foundational rights that are essential for a democratic society to exist and thrive. The right to effective remedy, enshrined in Article 13 of the Convention, both for victims of offences that took place either fully or partly online, as well as for those whose human rights were restricted by content moderation measures must be rigorously protected.

Due to ongoing evolution of crimes and technologies, prior review of self- or co-regulatory measures is not enough to ensure human rights are respected. Frequent review of the impact(s) of measures is also essential. The positive and negative obligations of states for the protection of human rights are equally applicable in the online environment.

3. Problem identification and targets

Content moderation is an effort to solve a problem. Therefore, it is crucial for the problem to be identified as clearly as possible, allowing for targeted solutions for varying problems. This is important to ensure necessity and proportionality.

The nature of the problem needs to be understood. Legislation that gives the burden of managing risk (which is, by definition, different for everyone) to internet intermediaries carries fundamentally different challenges compared with the removal of illegal content. Risk is, almost by definition, not

harm. [148] It is therefore crucial for the avoidance of unintended consequences, that any policy interventions that have the purpose of minimising risk are clearly recognised as such, in order to mitigate the particular problems of this approach, with the state taking its share of responsibility. They should also have clear targets, adjustment mechanisms and supervision, meaningful protection for freedom of expression, as well as tools to identify counterproductive impacts.

If the content moderation is being carried out in the context of a self- or co-regulatory scheme, there should also be mechanisms built in, to redesign, adapt or abandon the project, if either minimum standards are not achieved or if the nature of the problem evolves in a way which makes the identified approach not effective.

4. Meaningful decentralisation

As David Kaye, UN Special Rapporteur on Freedom of Opinion and Expression, detailed in his essay on this topic,[149] decentralisation is needed to moderate content in multinational or global contexts. Decentralised, multi-stakeholder, remunerated, empowered and independent moderation is essential to deal with problems on a regional level, taking regional peculiarities into account when dealing with the most difficult types of content.

> *Wherever the companies enjoy a market presence, they should develop multi-stakeholder councils, members of which they would compensate, to help them evaluate the hardest kinds of content problems, to evaluate emerging issues, and to dissent to the highest levels of company leadership.*[150]

Furthermore, his analysis is that clarity with regard to algorithmic decision-making would enable both individuals and academics to "register serious challenges" to enforcement of decisions. In the absence of readily available data, research and challenges to decisions become impossible. Consistent with the requirements for transparency above, adequate data needs to be made available to civil society and technical and academic researchers to facilitate ongoing analysis.

5. Communication with the user

Content moderation implies restriction of fundamental freedoms. These restrictions should respect human rights norms and be as transparent as possible towards the public, complainants, victims, and those whose content is removed.

a) Clarity and accessibility of terms of service

In addition to the full respect to the human rights of all stakeholders, all available tools should be used to ensure that the terms of service of an online intermediary are as clear and accessible as possible. The application of those rules should also be predictable, in line with human rights law. In line with David Kaye's analysis, human rights law has developed a language to define frameworks that can be used to articulate and ensure respect for democratic norms and counter authoritarian demands. It should be used.

b) Clarity on communication with users

Individuals who wish to complain about apparently illegal content, or content that apparently breaches the internal rules of an internet intermediary, should be given the tools to communicate

[148] Livingstone, Sonia; Kalmus, Veronika; Talves, Kairi (2014). "Girls' and boys' experiences of online risk and safety." In: Carter, Cynthia; Steiner, Linda; McLaughlin, Lisa (Ed.). The Routledge Companion to Media and Gender (190−200). London: Routledge, p 192.

[149] David Kaye, "A New Constitution for Content Moderation", June 25 2019, https://onezero.medium.com/a-new-constitution-for-content-moderation-6249af611bdf (last accessed 13 May 2020).

[150] Idem

their complaint to the company in the most specific way possible. Where relevant, internet intermediaries, in open dialogue with appropriate representative organisations, should ensure that their complaints mechanisms are victim sensitive.[151]

Those who post content should be given clear, balanced, and human rights-compatible rules that are implemented and enforced in a balanced, predictable way and not at the discretion of the platform. Those who access such content should also have easy access to these rules, and the right and the tools to make specific complaints, if they so wish.

Content should not be taken offline immediately, if it is not urgent that this be done. Instead, the individual who uploaded the content should be given clear information about why their content may have breached terms of service or the law, have the right to defend their upload within a set timeframe and, in any case, the right to a meaningful appeal.

Certain content does need to be taken offline as quickly as possible, due to the nature of the content or its impact on victims. Such content needs to be well defined and the process for reviewing it, deleting it and, as necessary, putting it back online, needs to be predictable, accountable and proportionate.

6. High level administrative safeguards

a) Clear legal and operational framework

A clear and predictable legal framework is essential to ensure that restrictions are provided for by instruments that are law or have the quality of law. States should ensure that terms of service are clear, balanced and equitably enforced. Laws on illegal content should be as clear and as harmonised as possible on an international level. States should ensure the existence of competent, independent authorities with the right to issue takedown orders. Finally, care should be taken to ensure that intermediaries are not unduly incentivised to restrict freedom of expression or other human rights. Appropriate and dissuasive redress rules should be put in place to discourage malicious reporting by individuals and to discourage over-compliance by internet intermediaries.

b) Supervision to ensure human rights compliance

Meaningful transparency on governance, decision-making processes, and details of how, when, why and how often, what content was removed, or not, and for what reason, can form the basis of meaningful measures to identify breaches of human rights. All data that is not personal should be made publicly available, on the basis of agreed industry standards on the methodology and format for such reporting. All such data should be made available to appropriate decentralised, multi-stakeholder, organisationally independent governance structures.

Particular safeguards are needed with regard to data protection in respect of content moderation. Care is needed when processing personal data of complainants and also of those accused of uploading illegal or unwelcome content (content that an internet intermediary may wish to restrict that is not illegal *per se*). This is especially the case when an individual is accused of posting potentially illegal information, where this content reveals sensitive personal data and/or where this leads to the processing of data that would otherwise not be necessary for the provision of the service being retained.

States should also take whatever action is needed to identify and prevent over-compliance and discrimination in content moderation.

[151] Alexander Brown, 2020, op cit, chapter VII.

c) *Evaluation and mitigation of "gaming" of complaints mechanisms*

Good transparency reporting will allow both companies and the public to be able to identify the gaming of companies' complaints mechanisms. This should be viewed as a priority and ongoing task for self- and co-regulatory schemes, particularly as available information points to such gaming being particularly harmful to women and minorities. This can be seen, for example, in the cases of "Women on Waves" and "Kick Out Zwarte Piet" in the Netherlands, which are described above.

d) *Ensuring consistency and independence of review mechanisms*

A crucial component to ensure the consistency and independence of a review mechanism is data. If enough data on decisions is made public and, where necessary, made available to independent third parties and if enough sample cases are made available to an independent and impartial body for proactive review, whose findings are meaningfully taken into account by the internet intermediary, then a high degree of consistency and independence can be assured.

Special attention must be given to ensuring that users feel empowered and listened to when appeals are made in relation to content that is, or is going to be, deleted. Similarly, transparency is needed to ensure that complainants are given adequate information to understand if and why their complaints have not led to the removal of the content in question.

e) *Recognising the human challenges of human content moderation*

In addition to the personnel management provisions of the Committee of Ministers Recommendation on the roles and responsibilities of internet intermediaries,[152] due attention must also be given to the labour rights and mental health of all workers involved in manual review of content which may be shocking, disturbing, or otherwise likely to have a psychological impact on the individuals concerned. This is particularly the case when internet intermediaries outsource this task to third parties, possibly based in other countries with different, and possibly less protective, labour laws.

f) *Ensuring protection of privacy and data protection*

Content moderation implies the processing of significant amounts of personal data related to the user, complainant and the nature of the content in question. It would be valuable for Council of Europe member States to ensure that adequate legal grounds exist in national law for this processing.

Internet intermediaries also need to take care that content moderation does not lead to inadvertent data protection breaches. For example, text or images used to replace removed content should avoid disclosure of sensitive data or, in the absence of a legal ruling, accusations of illegality.

g) *Victim redress*

Victims of illegal activity: As content moderation has the removal of content as its focus, states should additionally consider the rights of victims of illegal content, as a complement to such activities. This is necessary to ensure full support to victims to negate or mitigate the damage that has been caused.

Victims of unjustified takedowns: Appropriate measures are also needed to compensate victims of unjustified takedowns and to avoid such problems arising. Problems can be avoided by not removing content if this can be avoided, penalties for malicious reporting and by ensuring that internet intermediaries are not unduly incentivised to remove content. States should look beyond calculations based on financial loss. They should ensure that the moral cost of undue restrictions of freedom of expression, both for the direct victim and for society as a whole, is compensated.

[152] Committee of Ministers Recommendation (2018)2, op cit.

7. Addressing the peculiarities of self- and co-regulation in relation to content moderation

Traditional media self-regulation normally involves an entity regulating its own editorial decisions. Online self- and co-regulation is sometimes the same (when platforms regulate their own decisions for content that is entirely under their own control) and sometimes entirely different (when regulating the communication of their users, particularly at scale). As a result, assumptions based on experience of traditional media self-regulation are misleading in the context of most internet intermediary self-regulation. This fact should be actively considered in relation to any planned self- or co-regulatory scheme.

We saw above that different motivations and structures of self- and co-regulation have significant impacts on accountability and on the effectiveness of the measures in question. We also saw that different forms of self- and co-regulation imply different responsibilities for the State. It is crucial, therefore, both for compliance with human rights obligations and for the effectiveness of the measures being implemented, that the role of the state be honestly acknowledged, to ensure accountability and to build on experience of similar measures in this and other environments.