2018 10th International Conference on Cyber Conflict
CyCon X: Maximising Effects
T. Minárik, R. Jakschis, L. Lindström (Eds.)
2018 © NATO CCD COE Publications, Tallinn

Internet Intermediaries and Counter-Terrorism: Between Self-Regulation and Outsourcing Law Enforcement[1]

Krisztina Huszti-Orban
School of Law
University of Minnesota
Minneapolis, United States
khusztio@umn.edu

Abstract: Recent years have seen increasing pressure on Internet intermediaries that provide a platform for and curate third-party content to monitor and police, on behalf of the State, online content generated or disseminated by users. This trend is prominently motivated by the use of ICTs by terrorist groups as a tool for recruitment, financing, and planning operations. States and international organizations have long called for enhanced cooperation between the public and private sectors to aid efforts to counter terrorism and violent extremism. However, as the Special Rapporteur on Freedom of Expression noted in his latest report to the Human Rights Council, 'the intersection of State behaviour and corporate roles in the digital age remains somewhat new for many States'.

Detailed information on the means and modalities of content control exercised by online platforms is scarce. Terms of service and community standards are commonly drafted in terms that do not provide sufficiently clear guidance on the circumstances under which content may be blocked, removed or restricted, or access to a service may be restricted or terminated. Users have limited possibilities to challenge decisions to restrict material or access to a service. Moreover, as private bodies, such platforms are generally subject to limited democratic or independent oversight. At the same time, having private actors such as social media companies increasingly undertake traditionally public interest tasks in the context of Internet governance is likely unavoidable, as public authorities frequently lack the human or technical resources to satisfactorily perform these tasks.

1 This work was supported by the UK's Economic and Social Research Council [grant number ES/M010236/1].

Against this background, this paper aims to examine ways to define the contours of the division of responsibilities in countering terrorism and violent extremism between the public and private spheres. It addresses ways to ensure that Internet intermediaries carry out quasi-enforcement and quasi-adjudicative tasks in a manner compliant with international human rights norms and standards.

Keywords: *terrorism, violent extremism, human rights, Internet intermediaries, freedom of expression*

1. INTRODUCTION: ONLINE PLATFORMS AS GATEKEEPERS OF THIRD-PARTY CONTENT

It is difficult to overstate the role of the Internet intermediaries that provide a platform for and curate online content in facilitating the public's access to seek, receive, and impart information, including discourse on issues of public interest. Individuals' exercise of free speech is increasingly channelled through online platforms, which also enable governments to communicate with their constituencies and similarly facilitate the dissemination of messages by other actors. Many major online platforms (social media portals and search engines being prime examples) function on the basis of business models centred around hosting third-party content. The companies running these platforms regularly claim that the platforms function as mere distribution channels that exercise no or limited editorial intervention over the content published. Some of these sites have extremely high levels of user activity and interactivity,[2] allowing them to reach broad and diverse audiences in a manner that was not feasible before.[3] This, at the same time, makes meaningful real-time monitoring challenging or even impossible and editorial intervention time- and resource-intensive.

Online platforms regulate their use through terms of service and community standards. The private regulatory mechanisms used by these platforms generally represent an efficient alternative to public regulation in the online space. The terms and standards are pre-established and unilaterally imposed on all users who want access to the services offered, providing the platform with quasi-normative power when it comes to user behaviour. This power extends not only to the substantive aspects of use, such

2 It has been reported that every 60 seconds 510,000 comments are posted on Facebook, 293,000 statuses are updated, and 136,000 photos are uploaded. See Zephoria Digital Marketing, 'The Top 20 Valuable Facebook Statistics – Updated January 2018', 8 May 2017, https://zephoria.com/top-15-valuable-facebook-statistics/, accessed 15 January 2018. The daily video content watched on YouTube has reached 1billion hours this year. See YouTube Official Blog, 'You know what's cool? A billion hours' (27 February 2017) https://youtube.googleblog.com/2017/02/you-know-whats-cool-billion-hours.html, accessed 15 January 2018.

3 See Dave Chaffey, 'Global social media research summary 2017' (Smart Insights, 27 April 2017) http://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/, accessed 15 January 2018.

as the content that users are authorized to share or access, but also to enforcement-related ones, such as the criteria for decision-making and the technical tools used for the implementation of such decisions. In addition to these quasi-normative and quasi-executive functions, platforms frequently enjoy quasi-adjudicative power by requiring that disputes with users are settled via internal or other alternative dispute resolution or remedy mechanisms.

Such private 'sovereignty' should nonetheless be subject to public scrutiny to avoid arbitrary or abusive use of power. This is particularly important in light of some platforms undertaking functions traditionally catered for by the State. The argument that online platforms have become the digital age equivalent of public squares has been gaining traction in recent years.[4] Due to their reach, use, and level of interactivity, some of these platforms arguably play a public interest role. Studies show that people have increasingly been getting their news from social media.[5] Social media platforms have further been instrumental in disseminating information about political developments at home and abroad, humanitarian crises, and allegations of violations and abuses committed by States and Non-State actors.[6] In some countries or provinces, certain social media platforms are so dominant that to many inhabitants they represent the Internet itself.[7] As such, the information these inhabitants have access to online is restricted to whatever is available on these platforms. As offline information flows in these contexts are frequently restricted, social media platforms may constitute the main source of information, including of public interest information.

4 See Alissa Starzak, 'When the Internet (officially) became the public square' (Cloudfare, 21 June 2017) https://blog.cloudflare.com/internet-became-public-square/, accessed 15 January 2018; Ephrat Livni, 'The US Supreme Court just ruled that using social media is a constitutional right' (Quartz, 19 June 2017) https://qz.com/1009546/the-us-supreme-court-just-decided-access-to-facebook-twitter-or-snapchat-is-fundamental-to-free-speech/, accessed 15 January 2018.

5 See Jordan Crook, '62% of U.S. adults get their news from social media, says report' (TechCrunch, 26 May 2016) https://techcrunch.com/2016/05/26/most-people-get-their-news-from-social-media-says-report/, accessed 15 January 2018; Jane Wakefield, 'Social media "outstrips TV" as news source for young people', (BBC News, 15 June 2016) http://www.bbc.co.uk/news/uk-36528256, accessed 15 January 2018.

6 Christoph Koettl, 'Twitter to the rescue? How social media is transforming human rights monitoring', (Amnesty International USA, 20 February 2013) http://blog.amnestyusa.org/middle-east/twitter-to-the-rescue-how-social-media-is-transforming-human-rights-monitoring/, accessed 15 January 2018; Juliette Garside, 'Rioters' use of social media throws telecoms firms into spotlight' (The Guardian, 21 August 2011) https://www.theguardian.com/business/2011/aug/21/riots-throw-telecoms-firms-social-media-controls-into-spotlight, accessed 15 January 2018; Clay Shirky, 'The Political Power of Social Media: Technology, the public sphere and political change' Foreign Affairs (January/ February 2011) Vol. 90, No.1, 28-41.

7 See Megan Specia and Paul Mozur, 'A war of words puts Facebook at the center of Myanmar's Rohingya crisis' (The New York Times, 27 October 2017) https://www.nytimes.com/2017/10/27/world/asia/myanmar-government-facebook-rohingya.html?mtrref=www.google.com, accessed 12 March 2018; Casey Hynes, 'Internet use is on the rise in Myanmar, but better options are needed' (Forbes, 22 September 2017) https://www.forbes.com/sites/chynes/2017/09/22/Internet-use-is-on-the-rise-in-myanmar-but-better-options-are-needed/#1ef96e44448e, accessed 12 March 2018; Corynne McSherry, Jeremy Malcolm, Kit Walsh, 'Zero Rating: What it is and why you should care' (Electronic Frontier Foundation, 18 February 2016) https://www.eff.org/deeplinks/2016/02/zero-rating-what-it-is-why-you-should-care, accessed 12 March 2018.

The full picture needs to be considered in light of technological developments that have provided new means and modalities for controlling the content available online. Online platforms and those who provide and facilitate access to them have considerable power in shaping the information that gets disseminated; that is, they have *de facto* authority when it comes to regulating online content. As offline news consumption continues to decrease, particularly with younger demographics, these actors can exert significant influence over individuals' access to information, freedom of opinion, expression, and association, and over interlinked political and public interest processes.[8] The issue has figured prominently in recent discussions centring around the role of social media in influencing democratic, including electoral, processes.[9]

In addition to these regulatory functions, platforms have increasingly been undertaking policing and law enforcement functions traditionally considered to be State tasks. At times, such roles are delegated through law, as is the case of the German Network Enforcement Act.[10] However, platforms increasingly undertake such functions without their being formally delegated by state authorities, in an attempt to avoid liability or pre-empt State regulation.

This paper aims to examine the division of responsibilities between the public and private sphere in countering terrorism and violent extremism in a context where the 'playground' is privately owned and operated infrastructure, with uneven levels of State regulation. It addresses means and modalities to ensure that Internet intermediaries, with particular focus on social media platforms, carry out quasi-enforcement and quasi-adjudicative tasks in a manner compliant with international human rights norms and standards. The analysis will pay particular attention to relevant developments in European Union (EU) laws and policies and Member State practices.[11]

2. STATE TRENDS TO OUTSOURCE ONLINE (CONTENT) POLICING

Recent years have seen increasing pressure on Internet intermediaries that provide a platform for and curate third-party content to monitor and police, on behalf of the State,

8 Bruce Schneier, Data and Goliath (New York: W.W. Norton & Company, 2015), 114-116.

9 See, for example, Ryan Goodman and Justin Hendrix, ‚Facebook users have the right to know how they were exposed to Russian Propaganda' (Just Security, 23 October 2017) https://www.justsecurity.org/46171/facebook-users-right-to-know-exposed-russian-propaganda/, accessed 12 March 2018; Hannes Grassegger and Mikael Krogerus, 'The data that turned the world upside down' (Motherboard VICE, 27 January 2017) https://motherboard.vice.com/en_us/article/mg9vvn/how-our-likes-helped-trump-win, accessed 12 March 2018.

10 Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken [Netzwerksdurchsetzungsgesetz - NetzDG] (2017).

11 The reasons for choosing to demonstrate related issues by reference to the EU framework are the more detailed nature of EU regulation and its interpretation and also numerous current developments at the EU and Member State level. Many of the concerns raised are, however, valid beyond the EU.

online content that is generated or disseminated by users. This trend is prominently motivated by the use of ICTs and social media, in particular, by terrorist groups as a tool for recruitment, propaganda outreach, fundraising, and planning operations.[12] Discussions on the role and responsibilities of relevant online platforms in preventing and countering terrorism and violent extremism have intensified in the wake of recent attacks perpetrated by individuals linked to or inspired by ISIL.[13] Some policy-makers have expressed dissatisfaction with the efficiency of monitoring terrorist and violent extremist content and have warned platforms about the need to 'do more' if they want to avoid State intervention through binding regulation and sanctions.[14]

For its part, the tech industry has attempted to tackle the problems posed by terrorist or extremist third-party content through coordinated initiatives aimed at bolstering the efficiency of individually taken measures. Coordinated initiatives include the Global Internet Forum to Counter Terrorism,[15] the EU Internet Forum, bringing together EU entities, governments and technology companies,[16] the Code of Conduct on Countering Illegal Hate Speech Online,[17] and the Shared Industry Hash Database,[18] to name a few. Individually, companies have pledged to take further action to counter the use of their platforms for terrorist and other unlawful purposes by employing

12 See Brendan I. Koerner, 'Why ISIS is winning the social media war' (Wired, April 2016) https://www.wired.com/2016/03/isis-winning-social-media-war-heres-beat/, accessed 15 January 2018; David P. Fidler, 'Countering Islamic State exploitation of the Internet' (Council on Foreign Relations, 18 June 2015) https://www.cfr.org/report/countering-islamic-state-exploitation-internet, accessed 15 January 2018.

13 Andrew Sparrow, Alex Hern, 'Internet firms must do more to tackle online extremism, says No 10' (The Guardian, 24 March 2017) http://www.theguardian.com/media/2017/mar/24/internet-firms-must-do-more-to-tackle-online-extremism-no-10, accessed 15 January 2018; Jessica Elgot, 'May and Macron plan joint crackdown on online terror' (The Guardian, 12 June 2017) https://www.theguardian.com/politics/2017/jun/12/may-macron-online-terror-radicalisation, accessed 15 January 2018.

14 Amar Toor, 'France and the UK consider fining social media companies over terrorist content' (The Verge, 13 June 2017) https://www.theverge.com/2017/6/13/15790034/france-uk-social-media-fine-terrorism-may-macron, accessed 15 January 2018; Samuel Gibbs, 'Facebook and YouTube face tough new laws on extremist and explicit video' (The Guardian, 24 May 2017) https://www.theguardian.com/technology/2017/may/24/facebook-youtube-tough-new-laws-extremist-explicit-video-europe, accessed 15 January 2018; Kate McCann, 'Facebook "must pay to police internet" or face fines: UK Parliament' (The Canberra Times, 30 April 2017) http://www.canberratimes.com.au/technology/technology-news/facebook-must-pay-to-police-internet-20170430-gvvz2e.html, accessed 15 January 2018.

15 Microsoft Corporate Blogs, 'Facebook, Microsoft, Twitter and YouTube announce formation of the Global Internet Forum to Counter Terrorism' (26 June 2017) https://blogs.microsoft.com/on-the-issues/2017/06/26/facebook-microsoft-twitter-youtube-announce-formation-global-internet-forum-counter-terrorism/, accessed 15 January 2018.

16 European Commission, 'EU Internet Forum: a major step forward in curbing terrorist content on the internet. Press release' (Brussels, 8 December 2016) http://europa.eu/rapid/press-release_IP-16-4328_en.htm, accessed 15 January 2018.

17 The initiative is from the European Commission, together with Facebook, Microsoft, Twitter and YouTube. The Code of Conduct is available at http://ec.europa.eu/justice/fundamental-rights/files/hate_speech_code_of_conduct_en.pdf, accessed 15 January 2018.

18 Google, 'Partnering to help curb the spread of terrorist content online' (5 December 2016) https://www.blog.google/topics/google-europe/partnering-help-curb-spread-terrorist-content-online/, accessed 15 January 2018.

artificial intelligence and 'human expertise' to identify 'extremist and terrorism-related' content.[19]

3. ONLINE PLATFORMS AND COUNTER-TERRORISM

Relevant corporate obligations are included in a variety of laws adopted at the national level, among others those tackling hate speech, cybercrime, counter-terrorism, violent extremism, and intermediary liability. Many jurisdictions also encourage self- and co-regulation.

A. Terrorism and Violent Extremism: Dilemmas of Definition

Despite a plethora of multilateral treaties, Security Council resolutions, and other international and regional instruments addressing terrorism-related issues,[20] an internationally agreed definition of terrorism or an agreed list of terrorism-related offences is lacking. As a result, relevant definitions are to be found in laws and policies adopted at the level of States, causing considerable discrepancies between different domestic frameworks.

Particularly pertinent to our context are preparatory and ancillary offences and, newly, offences criminalizing the advocacy of terrorism, including 'glorification', 'apology', 'praise' or 'justification' of terrorism.[21] United Nations human rights mechanisms and other stakeholders have raised concerns over some definitions lacking precision, stressing the potential negative human rights implications of definitions of terrorism

19 See, for example, Google, 'Four steps we're taking today to fight terrorism online' 18 June (2017) https://www.blog.google/topics/google-europe/four-steps-were-taking-today-fight-online-terror/, accessed 15 January 2018; Monika Bickert, Brian Fishman, 'Hard Questions: How We Counter Terrorism', (15 June 2017) https://newsroom.fb.com/news/2017/06/how-we-counter-terrorism/, accessed 15 January 2018; Twitter Inc. 'An update on our efforts to combat violent extremism' (18 August 2016) https://blog.twitter.com/official/en_us/a/2016/an-update-on-our-efforts-to-combat-violent-extremism.html, accessed 15 January 2018.

20 See United Nations Counter-Terrorism Implementation Task Force, International Legal Instruments, https://www.un.org/counterterrorism/ctitf/en/international-legal-instruments, accessed 15 January 2018.

21 The UN Human Rights Committee has stressed that offences such as 'praising', 'glorifying', or 'justifying' terrorism must be clearly defined to ensure that they do not lead to unnecessary or disproportionate interference with freedom of expression. See United Nations Human Rights Committee, General Comment 34. Article 19: Freedoms of opinion and expression (CCPR/C/GC/34), para. 46. Similarly, the Secretary-General and the UN Special Rapporteur on Counter-Terrorism have expressed concerns about the 'troubling trend' of criminalizing the glorification of terrorism, stating that this amounts to an inappropriate restriction on expression. See Protecting Human Rights and Fundamental Freedoms While Countering Terrorism. Report of the Secretary-General (A/63/337) and United Report of the Special Rapporteur on the Promotion and Protection of Human Rights and Fundamental Freedoms While Countering Terrorism (A/HRC/31/65).

and related offences that are overly-broad[22] or attach criminal sanctions to conduct that falls short of incitement to terrorism or advocacy of national, racial or religious hatred constituting incitement to violence.[23]

Laws and policies addressing violent extremism similarly raise definitional concerns. While the term 'violent extremism' and related notions such as 'extremism' and 'radicalization' are prominently present in current political discourse at the international, regional, and national levels, none of these terms have internationally agreed definitions.[24] Many of the relevant definitions found in domestic laws and policies have been criticized for being vague and at times encompassing manifestations that are lawful under international human rights law.[25] In some jurisdictions, these concepts have become dissociated from violence,[26] thereby raising the potential for abusive implementation, as such definitions risk selectively blurring the distinction between belief and violent conduct. Such approaches, especially when not accompanied by robust safeguards, risk leading to the suppression of views that deviate from the social norms accepted by the majority, under the guise of preventing extremism; and measures may target thought, belief, and opinion, rather than actual conduct.

22 See, for example, Protecting Human Rights and Fundamental Freedoms While Countering Terrorism. Report of the Secretary-General, (A/68/298); Report of the United Nations High Commissioner for Human Rights on the Protection of Human Rights and Fundamental Freedoms While Countering Terrorism (A/HRC/28/28); International Commission of Jurists, Report of the Eminent Jurists Panel on Terrorism, Counter-Terrorism and Human Rights (2009). See also, Cathal Sheerin, 'The threat of 'glorifying terrorism' laws' (IFEX, 2 February 2017) https://www.ifex.org/europe_central_asia/2017/02/02/glorifying_terrorism_charges/, accessed 12 March 2018; EDRi, 'European Union Directive on counter-terrorism is seriously flawed' (30 November 2016) https://edri.org/european-union-directive-counterterrorism-seriously-flawed/, accessed 12 March 2018; Amnesty International, 'EU: Orwellian counter-terrorism laws stripping rights under guise of defending them' (17 January 2017) https://www.amnesty.org/en/latest/news/2017/01/eu-orwellian-counter-terrorism-laws-stripping-rights-under-guise-of-defending-them/, accessed 12 March 2018; Amar Toor, 'France extends draconian anti-terrorism laws' (The Verge, 17 February 2016) https://www.theverge.com/2016/2/17/11031006/france-extends-state-of-emergency-paris-attacks , accessed 12 March 2018; Amnesty International, 'Tweet… if you dare. How counter-terrorism laws restrict freedom of expression in Spain' (March 2018), Index no. EUR 41/7924/2018.

23 See Article 20, International Covenant on Civil and Political Rights. See also, Amnesty International, 'Tweet… if you dare. How counter-terrorism laws restrict freedom of expression in Spain'. In France, the Constitutional Court has recently struck down an amendment to the Penal Code criminalizing 'regular consultation' of content deemed to be inciting or glorifying terrorism. See Nadim Houry, 'French legislators rebuked for seeking to criminalize online browsing' (Human Rights Watch, 15 December 2017) https://www.hrw.org/news/2017/12/15/french-legislators-rebuked-seeking-criminalize-online-browsing, accessed 12 March 2018; Conseil constitutionnel, Décision n° 2017-682 QPC du 15 décembre 2017.

24 Acknowledging this shortcoming, the Secretary-General in his Plan of Action to Prevent Violent Extremism stated that violent extremism is to be defined at the national level, while emphasizing that such definitions must be consistent with obligations under international human rights law.

25 See Report of the Special Rapporteur on the Promotion and Protection of Human Rights and Fundamental Freedoms While Countering Terrorism (A/HRC/31/65) and Report on Best Practices and Lessons Learned on How Protecting and Promoting Human Rights Contribute to Preventing and Countering Violent Extremism. Report of the United Nations High Commissioner for Human Rights (A/HRC/33/29).

26 A number of countries also target 'extremism' that is non-violent. For example, extremism is defined in the United Kingdom as 'vocal or active opposition to fundamental values, including democracy, the rule of law, individual liberty and the mutual respect and tolerance of different faiths and beliefs'. See HM Government, Prevent Strategy (2011), Annex A; HM Government, Counter-Extremism Strategy (2015, October), para. 1.

The potential and actual uses of the counter-terrorism and preventing violent extremism framework to stifle dissent, to persecute human rights defenders, journalists, and the political opposition, and to criminalize the work of humanitarian organizations has been addressed at length elsewhere.[27] Online platforms having to operationalize such laws and policies may find themselves contributing to the negative human rights impact of these frameworks. Even in cases where related domestic legal and policy frameworks do not present these shortcomings, the discrepancies between different domestic frameworks inevitably raise difficulties for online platforms, in particular those that operate worldwide (or at least in numerous jurisdictions), making it difficult to comply with all relevant domestic laws.

B. Online Platforms as De Facto Content Regulators

1) Means and Modalities of Content Review

Many platforms rely on a combination of artificial intelligence (AI) and human expertise to review and moderate content. The use of AI to spot terrorist or violent extremist content is a relatively new development,[28] and platforms such as Facebook acknowledge that it is a tool that must be complemented by human review.[29] Using algorithms to assess compliance with the law and terms of service or community standards provides for a time-efficient way for dealing with large volumes of material. It is one advocated by bodies such as the European Commission, which encourages online platforms to 'step up investment in, and use of, automatic detection technologies'.[30]

Algorithms, however, are not fool-proof, as they are not necessarily well-equipped to understand context, different forms of humour and satire,[31] and may not pick up on certain subtleties.[32] For example, hash-matching or even fingerprinting algorithms are not capable of analysing meaning or context, such as whether certain content contains

27 See Interagency Standing Committee, Sanctions Assessment Handbook: Assessing the Humanitarian Implications of Sanctions (United Nations, 2004); Kate Mackintosh and Patrick Duplat, 'Study of the Impact of Donor Counter-Terrorism Measures on Principled Humanitarian Action' (United Nations Office for the Coordination of Humanitarian Affairs and the Norwegian Refugee Council, July 2013).

28 See Monika Bickert and Brian Fishman, note 19.

29 Monika Bickert and Brian Fishman, 'Hard Questions: Are We Winning the War on Terrorism Online?' (Facebook, 28 November 2017) https://newsroom.fb.com/news/2017/11/hard-questions-are-we-winning-the-war-on-terrorism-online/, accessed 15 January 2018; Lynsey Barber, 'Facebook's now using artificial intelligence to remove terror content' (CityA.M., 29 November 2017) http://www.cityam.com/276626/facebooks-now-using-artificial-intelligence-remove-terror, accessed 15 January 2018.

30 European Commission, 'Communication on tackling illegal content online, towards enhanced responsibility of online platforms' (28 September 2017) https://ec.europa.eu/digital-single-market/en/news/communication-tackling-illegal-content-online-towards-enhanced-responsibility-online-platforms, accessed 15 January 2018.

31 Julia Krüger, 'Kommentar: Das Recht auf den Tweet' (Netzpolitik.org, 6 January 2018) https://netzpolitik.org/2018/kommentar-das-recht-auf-den-tweet/, accessed 15 January 2018.

32 See, for example Julia Reda, 'When filters fail: These cases show we can't trust algorithms to clean up the Internet' (28 September 2017) https://juliareda.eu/2017/09/when-filters-fail/, accessed 15 January 2018.

terrorist propaganda or hate speech, or reveals criminal intent.[33] As a result, they may end up removing not only videos produced by terrorist groups for recruitment purposes, but also media analysis of these videos, or even footage uploaded by human rights groups reporting on abuses.[34] Some machine-learning algorithms, such as natural language processing tools, are better suited for the kind of analysis required in this context. However, even their use comes with limitations. Experts argue that these tools cannot be applied with the same reliability across different contexts, as language use differs across different cultural, demographic, and linguistic groups.[35] An algorithm trained to parse out anti-Muslim hate speech may achieve lower levels of accuracy when attempting to identify anti-Semitic hate speech, for example. As with any machine learning algorithm, these tools can also amplify existing biases (including social and other bias existing in a language). This may result in algorithms over-censoring groups that are already marginalized.[36] Dialects that are underrepresented in mainstream text are also more likely to be misinterpreted, leading to algorithms performing less accurately,[37] and many of the existing natural language processing tools only work for English or other high-resource languages.[38]

These limitations suggest that unchecked use of algorithms for content management may lead to screening that is over- or under-inclusive. The margin of error would prove particularly problematic in the case of large online platforms. For example, Facebook has at some point reported that it receives one million user violation reports a day.[39] If all these reports were processed through AI tools, it would mean hundreds of thousands of mistaken decisions per day.[40] For meaningful oversight of decisions made by AI tools, integrating the human-in-the-loop principle needs to be ensured. Unfortunately, most social media platforms do not provide meaningful information

33 *Ibid*. See also Evan Engstrom and Nick Feamster, 'The limits of filtering: A look at the functionality & shortcomings of content detection tools' (Engine, March 2017) 13-15 and 17-21.

34 See, for example, Daphne Keller, 'Problems with filters in the European Commission's platforms proposal' (Stanford Center for Internet and Society, 5 October 2017) http://cyberlaw.stanford.edu/blog/2017/10/problems-filters-european-commissions-platforms-proposal, accessed 12 March 2018.

35 See Bermingham et al., 'Combining Social Network Analysis and Sentiment Analysis to Explore the Potential for Online Radicalisation', Proceedings of the International Conference on Advances in Social Network Analysis and Mining (2009), 3; Su Lin Blodgett and Brendan O'Connor, 'Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English', Proceedings of the Fairness, Accountability, and Transparency in Machine Learning Conference (2017) 1-2, https://arxiv.org/pdf/1707.00061.pdf, accessed 15 January 2018.

36 Jieyo Zhao et al., 'Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-Level Constraints', Proceedings of the Conference on Empirical Methods in Natural Language Processing (2017), https://arxiv.org/pdf/1707.09457, accessed 15 January 2018.

37 Su Lin Blodgett and Brendan O'Connor, note 35, 1-2; Rachael Tatman, 'Gender and Dialect Bias in YouTube's Automatic Captions', Proceedings of the First Association for Computational Linguistics Workshop on Ethics in Natural Language Processing, 53–59 (2017), http://rachaeltatman.com/sites/default/files/papers/EthNLP06.pdf, accessed 15 January 2018. See also Natasha Duarte, Emma Llanso, Anna Loup, 'Mixed Messages? The Limits of Automated Social Media Content Analysis' (Centre for Democracy and Technology, November 2017) 15.

38 *Id*., 14.

39 Sara Ashley O'Brien, 'Facebook gets 1 million user violation reports a day' (CNN Tech, 12 March 2016) http://money.cnn.com/2016/03/12/technology/sxsw-2016-facebook-online-harassment/index.html, accessed 15 January 2018.

40 Natural language processing tools reportedly do not possess an accuracy rate higher than 80%. See Natasha Duarte, note 37, 5.

on content review procedures and the criteria that determine whether certain content will be reviewed by AI, human moderators, or both.[41]

Having content reviewed by human moderators does not necessarily assuage all concerns. Assessing what may amount to hate speech, incitement to terrorism, 'glorification' of terrorism or violent extremist content frequently requires a rather complex analysis to be conducted by a highly trained, specialized, and adequately resourced workforce. The reality, however, does not seem to fit this picture. Reports indicate that low-paid and insufficiently trained moderators frequently end up being the *de facto* gatekeepers of freedom of expression online.[42] Moreover, bearing in mind the overwhelming pace at which content is posted, relying primarily on human monitoring, particularly in near real-time, would be next to impossible.

Many large social media platforms operate worldwide, or at least in numerous jurisdictions. This makes it difficult or even impossible to produce a universally suitable set of rules for their algorithms and moderators. As described above, such rules need to take into account the differences between domestic legal systems and the scope of prohibited content in different jurisdictions and linguistic, cultural, social, and other contexts.

2) Safeguards, Transparency, and Accountability

Detailed information on the means and modalities of content control exercised by online platforms is scarce. Terms of service and community standards are commonly drafted in vague terms and do not provide sufficiently clear guidance on the circumstances under which content may be blocked, removed or restricted, or access to a service restricted or terminated, including the criteria used for such assessments. Facebook's Director of Global Policy Management, Monika Bickert, explained that the company does not share details of its policies to avoid encouraging people 'to find workarounds'.[43] This also means reduced transparency, including when it comes to the internal consistency of the application of these policies, and may as a result lead to reduced accountability.

41 See Monika Bickert, note 19. While the so-called 'Facebook files' provide some insight into the moderation process, many questions remain. Moreover, moderation policies of other major social network platforms remain obscure.

42 See Olivia Solon, 'Counter-terrorism was never meant to be Silicon Valley's job. Is that why it's failing?' (The Guardian, 29 June 2017) https://www.theguardian.com/technology/2017/jun/29/silicon-valley-counter-terrorism-facebook-twitter-youtube-google; accessed 15 January 2018; Olivia Solon, 'Underpaid and overburdened: The life of a Facebook moderator' (The Guardian, 25 May 2017) https://www.theguardian.com/news/2017/may/25/facebook-moderator-underpaid-overburdened-extreme-content, accessed 15 January 2018; Till Krause and Hannes Grassegger, 'Inside Facebook' (Süddeutsche Zeitung, 15 December 2016) http://www.sueddeutsche.de/digital/exklusive-sz-magazin-recherche-inside-facebook-1.3297138, accessed 15 January 2018; Nick Hopkins, 'Facebook struggles with 'mission impossible' to stop online extremism' (The Guardian, 24 May 2017) https://www.theguardian.com/news/2017/may/24/facebook-struggles-with-mission-impossible-to-stop-online-extremism, accessed 15 January 2018.

43 Monika Bickert, 'At Facebook we get things wrong – but we take our safety role seriously' (The Guardian, 22 May 2017) https://www.theguardian.com/commentisfree/2017/may/22/facebook-get-things-wrong-but-safety-role-seriously, accessed 15 January 2018.

Information provided *ex post facto* is similarly lacking. Users are frequently not informed of the origin of removal requests, the procedure that led to removal or rejection of removal and the criteria used.[44] They also have limited possibilities to challenge decisions to restrict content or access to a service. To tackle this shortcoming, the recently adopted German Network Enforcement Act requires companies to report on a biannual basis describing their means and modalities for handling complaints and disclosing the criteria for removing or blocking content. It similarly calls on companies to inform both the complainant and the users affected by particular measures, including the reasoning for the decision. The law, however, does not explicitly require companies to provide users with the option to challenge these decisions.

As relevant measures by private companies are generally taken in enforcement of terms of service and not on the basis of specific legislation, it is frequently not possible to challenge them in court. Platforms may also impose internal or other alternative dispute resolution mechanisms, should disputes arise. Moreover, as private bodies, such platforms are generally not subject to democratic or independent oversight in the way that public authorities are, despite their effectively carrying out regulatory, executive, and adjudicative functions.[45] Removing the possibility of independent, including judicial, review of measures that interfere with human rights is problematic in general and particularly so having in mind recent legal and policy developments. Businesses are potentially facing fines and sanctions imposed by States if they do not restrict unlawful content.[46] On the other hand, should they remove lawful content in the process, affected individuals have limited avenues of redress. In case of doubt, businesses will more likely err on the side of over-censoring.

C. The Scope of Responsibility of Online Intermediaries

Online platforms that host or store user-generated content and enable access to and retrieval of this content by the author and other users[47] qualify as Internet intermediaries. Such intermediaries, as opposed to authors and publishers of content, are generally protected against liability for third-party content, with certain caveats. The scope of this exemption differs depending on jurisdiction.[48] For example, under the EU e-Commerce Directive, hosting intermediaries do not incur liability as long

44 See Annemarie Bridy and Daphne Keller, 'U.S. Copyright Office Section 512 Study: Comments in Response to Notice of Inquiry' (31 March 2016) 29.

45 'Zachary Loeb – Who moderates the moderators? The Facebook files' (Boundary 2, 7 June 2017) http://www.boundary2.org/2017/06/zachary-loeb-who-moderates-the-moderators-on-the-facebook-files/, accessed 15 January 2018.

46 See Section C *infra* : The Scope of Responsibility of Online Intermediaries.

47 Monica Horten, 'Content 'responsibility': The looming cloud of uncertainty for Internet intermediaries' Center for Democracy and Technology (September 2016) 5. See also Jaani Riordan, The Liability of Internet Intermediaries (Oxford University Press, 2016) Chapter 2.

48 See Article 19, 'Internet Intermediaries: Dilemma of Liability' (2013); Eric Goldman, 'Facebook isn't liable for fake user account containing non-consensual pornography' (Forbes, 8 March 2016) https://www.forbes.com/sites/ericgoldman/2016/03/08/facebook-isnt-liable-for-fake-user-account-containing-non-consensual-pornography/#40ba670379b2, accessed 15 January 2018.

as they 'expeditiously' remove or disable access to illegal content once they have 'actual knowledge' of its existence.[49] Under EU law, it is not permitted to impose a general obligation to monitor content or to 'actively seek facts or circumstances indicating illegal activity'.[50] Similarly, so-called 'notice and stay-down' injunctions, involving an obligation to ensure that content, once removed, will not reappear on the platform, are also problematic to the extent that their implementation requires general monitoring.

The idea of introducing such a burden on intermediaries has emerged in current debates, with policy-makers calling for stricter regulation of the liability of Internet intermediaries when it comes to countering terrorism, violent extremism, and hate speech. Proposals include imposing fines and other sanctions on social media platforms 'that fail to take action against terrorist propaganda and violent content',[51] and even having social media companies bear the costs of authorities policing content online.[52] The introduction of criminal liability for platforms was discussed and ultimately rejected by the European Parliament in the context of the Directive on Combating Terrorism. However, the European Commission, in its *Communication on Tackling Illegal Content Online: Towards enhanced responsibility of online platforms*, recommended that tech companies proactively look to identify illegal content on their platforms with the help of artificial intelligence, stressing that 'online platforms should also be able to take swift decisions […] without being required to do so on the basis of a court order or administrative decision'.[53] The Commission considers that online platforms can take the recommended proactive measures without fear of losing their liability exemption under the e-Commerce Directive.[54]

Other developments similarly come close to recommending or requiring proactive monitoring by intermediaries, potentially also affecting the internal consistency of the EU legal framework. Article 28a of the review proposal to the Audio-Visual Media Services (AVMS) Directive[55] provides that video-sharing platforms[56] must

49 Article 14, Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (e-Commerce Directive).

50 Article 15, e-Commerce Directive. See also *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v. Netlog NV*. Case C-360/10 (2012) (European Court of Justice).

51 Toor, note 14; Gibbs, note 14.

52 See House of Commons Home Affairs Committee, 'Hate crime: abuse, hate and extremism online' (25 April 2017); McCann, note 14.

53 European Commission, note 30.

54 While the Communication addresses the compatibility of such proactive measures with Article 14 of the e-Commerce Directive, it does not pay similar attention to Article 15.

55 European Commission, Proposal for a Directive of the European Parliament and of the Council Amending Directive 2010/13/EU: On the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audio-visual media services in view of changing market realities.

56 It must be noted that some civil society organizations and some Member States caution against the inclusion of video-sharing platforms, in particular social media ones, within the scope of the Directive. See EDRi, 'EDRi Position on AVMSD Trilogue Negotiations' (14 September 2017) https://edri.org/files/AVMSD/edriposition_trilogues_20170914.pdf, accessed 15 January 2018.

take measures to 'protect all citizens' from content containing incitement to violence, discrimination or hate.[57] In addition to providing for a rather vague definition of such content,[58] the proposed provision may be interpreted as requiring proactive monitoring.[59]

As a result of such developments, the EU will have to assess the compatibility of the e-Commerce Directive with other instruments addressing the role of Internet intermediaries in combating hate speech and other illegal content, especially in light of the decision not to reopen the e-Commerce Directive. It is in this vein that the European Commission has adopted the above-mentioned *Communication on Tackling Illegal Content Online*[60] and is developing measures that set common requirements across the Union for companies when it comes to removing illegal content, as a means to avoid 'overzealous rules that differ between EU countries'.[61]

What seems to be missing is the human rights-based analysis of such new obligations. This shortcoming comes even though human rights concerns posed by far-reaching intermediary liability and, in particular, its negative impact on freedom of speech and interlinked rights, have repeatedly been flagged by international human rights mechanisms[62] and civil society actors.[63] It is questionable whether the course of action proposed in the Commission's *Communication* can be construed in line with human rights standards,[64] including as spelled out in the EU Council's *Human Rights*

57 See EDRi, 'EDRi's analysis on the CULT compromise on Article 28a of the draft Audiovisual Media Services Directive (AVMSD) proposal' (13 April 2017) https://edri.org/files/AVMSD/compromise_article28a_analysis_20170413.pdf, accessed 15 January 2018.

58 For example, a compromise amendment under discussion provides for the following: 'protect all citizens from content containing incitement undermining human dignity, incitement to terrorism or content containing incitement to violence or hatred directed against a person or a group of persons defined by reference to nationality, sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age, gender, gender expression, gender identity, sexual orientation, residence status or health.' (emphasis added) See European Parliament. Committee on Civil Liberties, Justice and Home Affairs. (2016) Amendments 47-171. (2016/0151(COD)).

59 While the draft explicitly mentions that it is without prejudice to articles 14 and 15 of the e-Commerce Directive, the intended scope of the duty of care is still unclear. See also Horten, note 47, 14.

60 European Commission, 'Liability of online intermediaries', (15 June 2017) https://ec.europa.eu/digital-single-market/en/liability-online-intermediaries, accessed 15 January 2018.

61 Catherine Stupp, 'Gabriel to start EU expert group on fake news' (Euractiv, 30 August 2017) https://www.euractiv.com/section/digital/news/gabriel-to-start-eu-expert-group-on-fake-news/, accessed 15 January 2018.

62 See Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (A/HRC/35/22), para. 49. See also, Joint Declaration by the United Nations 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, presented at the UNESCO World Press Freedom Day event (3 May 2016).

63 See note 48.

64 For criticism of the Communication, see for example Daphne Keller, note 34; Graham Smith, 'Towards a filtered internet: the European Commission's automated prior restraint machine' (Cyberleagle, 25 October 2017) http://www.cyberleagle.com/2017/10/towards-filtered-internet-european.html, accessed 12 March 2018.

Guidelines on Freedom of Expression Online and Offline,[65] bearing in mind its emphasis on protecting intermediaries from an obligation of blocking Internet content 'without prior due process'. Indeed, the *Communication* seems to stress *ex post facto* modalities of redress at the expense of 'prior due process'. In this respect, it states that platforms should be able to take 'swift decisions' to take action with respect to illegal content 'without being required to do so on the basis of a court order or administrative decision'. This is the case in particular when such content has been flagged by a law enforcement authority. Law enforcement authorities may be so-called 'trusted flaggers', together with other 'specialized entities with specific expertise in identifying illegal content'. In some cases, platforms 'may remove content upon notification from the trusted flagger without further verifying the legality of the content themselves'.

One entity identified as a trusted flagger in the context of assessing terrorist and violent extremist content is the Internet Referral Unit (IRU) of Europol. The IRU flags content that contravenes the EU legal framework related to terrorism and also content that goes against the terms of service set by platforms.[66] However, terms of service instituted by platforms commonly impose restrictions that go beyond what could lawfully be imposed in compliance with freedom of expression standards.[67] This approach creates the risk that content will be blocked, filtered or removed beyond what would be permissible under international human rights law. It may also result in undermining regular safeguards that protect against excessive interference, including the right to an effective remedy, as the end decision is ultimately delegated to private entities.[68]

Relevant developments have to be noted at Member State level as well. Germany has recently adopted the controversial[69] Network Enforcement Act,[70] imposing onerous obligations on social media platforms with more than two million registered users. Platforms falling within the ambit of the law face fines of up to €50 million if they

65 Council of the European Union, EU Human Rights Guidelines on Freedom of Expression Online and Offline (Foreign Affairs Council meeting, Brussels, 12 May 2014) http://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/EN/foraff/142549.pdf, accessed 12 March 2018.

66 See European Parliament, 'Question for written answer to the Commission' (16 March 2017) http://www.europarl.europa.eu/sides/getDoc.do?type=WQ&reference=E-2017-001772&language=FR, accessed 12 March 2018; Answer given by Mr Avramopoulos on behalf of the Commission (12 June 2017) http://www.europarl.europa.eu/sides/getAllAnswers.do?reference=E-2017-001772&language=EN, accessed 12 March 2018. See also, Graham Smith, note 64.

67 See e.g. Elizabeth Nolan Brown, 'YouTube says no to sexual humor, profanity, partial nudity, political conflict, and "sensitive subjects" in partner content' (Reason, 1 September 2016) http://reason.com/blog/2016/09/01/youtube-bans-sex-drugs-and-politics , accessed 12 March 2018. As privately-run outlets, social media platforms can of course decide to shape the content hosted by them in order to facilitate the creation of a space that fits their business model, by enabling a more family-friendly or minor-friendly environment, for example.

68 See European Digital Rights (EDRi). (2011, January). The Slide from 'Self-Regulation' to 'Corporate Censorship'. Retrieved from https://edri.org/files/EDRI_selfreg_final_20110124.pdf.

69 'Wirtschaft und Aktivisten verbünden sich gegen Maas-Gesetz' (Der Spiegel, 11 April 2017) http://www.spiegel.de/netzwelt/netzpolitik/heiko-maas-wirtschaft-und-netzszene-protestieren-gegen-hassrede-gesetz-a-1142861.html, accessed 15 January 2018.

70 Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken.

fail to remove or block access to 'clearly illegal' content within 24 hours[71] and other illegal content within 7 days[72] after having been put on notice through a complaint. The law includes no guidance on how to distinguish 'clearly illegal' entries from merely 'illegal' ones. Such lack of clear guidance, when paired with a threat of hefty fines, becomes a definite incentive to over-censor in case of doubt.

Implementation of the Act started on the 1st of January 2018 and related incidents have already drawn attention to the limits of algorithmic moderation[73] as well as the discrepancies in the approach to moderating content demonstrated by different social media platforms.[74] In addition to cases of lawful content being removed by overeager platforms, some argue that it also results in obstructing prosecution of related crimes, as deletion of online content frequently results in deletion or improper retention of evidence needed to plead the case in court.[75] The Act will inevitably influence how major social media sites approach users' freedom of expression, with its impact in all probability extending beyond Germany's borders due to the cross-border nature of information flows and also the likelihood of it influencing similar legal and policy initiatives in other jurisdictions.

Changes in laws and policies aimed at more effectively tackling terrorist and extremist content and hate speech have also been contemplated in other jurisdictions. In this respect, the UK House of Commons Home Affairs Committee has recommended that Internet intermediaries proactively identify illegal content and expressed dissatisfaction with such platforms for only reviewing content after it has been flagged by users or other stakeholders and for not ensuring that blocked and removed content does not resurface.[76] Similarly, the *French-British Action Plan on the Use of the Internet for Terrorism Purposes*[77] (also known as the Macron-May Plan) calls on platforms to proactively identify terrorist content and prevent it from being made available by automating the detection and suspension or removal of content, based on both the posting person or entity and the actual content of the post. This measure

71 Unless the social media network agrees a different timeline with the competent law enforcement authority. Netzwerksdurchsetzungsgesetz, Article 1 §3 (2) No. 2.

72 Unless the unlawful character of the content in question depends on factual circumstances to be determined or unless the social media network transmits the case to an authorized self-regulatory mechanism (Einrichtung der regulierten Selbstregulierung). Netzwerkdurchsetzungsgesetz, Article 1 §3 (2) No. 3.

73 See note 31.

74 *Ibid*.

75 Bernhard Rohleder, 'Germany set out to delete hate speech online. Instead, it made things worse.' (The Washington Post, 20 February 2018) https://www.washingtonpost.com/news/theworldpost/wp/2018/02/20/netzdg/?utm_term=.331d14c7fb0a, accessed 12 March 2018.

76 House of Commons Home Affairs Committee, note 52. See also, Elliot Harmin, '"Notice-and-stay-down" is really "filter-everything"' (Electronic Frontier Foundation, 21 January 2016) https://www.eff.org/deeplinks/2016/01/notice-and-stay-down-really-filter-everything, accessed 15 January 2018.

77 French-British Action Plan on the Use of the Internet for Terrorism Purposes (Paris, 13 June 2017) https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/619333/french_british_action_plan_paris_13_june_2017.pdf, accessed 15 January 2018.

drew criticism for advocating both far-reaching monitoring and prior restraint.[78] The Plan also recommends measures that go beyond the existing 'notice and take-down' process, which has also been reinforced through the establishment of Europol's Internet Referral Unit and the UK's own domestic referral unit, raising the possibility of a 'notice and stay-down' obligation.

4. CONCLUSION

The Internet has frequently been described as a neutral tool that can be instrumentalised in various ways.[79] It is fundamental in facilitating the public's ability to seek, receive, and impart information and may provide a platform for persons and groups that are less included in debates of public interest, such as women and individuals belonging to minority groups, but it also enables terrorist groups and other criminal actors to convey their messages and use it as a recruitment and operational planning tool.

As online content continues to be generated at a staggering rate, attempts to control its flow encounter significant challenges and, due to the particularities of the digital space, tech companies running these online platforms are better positioned to regulate their functioning, while State powers in this respect may be more limited. There are clear expectations on the part of States that online platforms take more responsibility when it comes to illegal third-party content. Many governments view the use of automated decision-making tools as an essential component of handling content. The choice is understandable, having in mind the volume of the material that is being produced, the pace of such production and the need to take swift action. However, the limitations of existing technology are significant. If algorithms are used for regulating content, they become the rule, the rule-maker in the case of machine learning algorithms, and the tool for enforcement. The rules behind the algorithms become the *de facto* standards for the platform and beyond.

The duty of States to protect the human rights of those within their jurisdiction, including from undue interference by third parties such as businesses, is well-established. Outsourcing such tasks – whether formally or informally, through actively encouraging corporate governance or through omission or acquiescence – without establishing adequate safeguards and oversight systems, fails to comply with that duty.[80] The rise of automated processes without a corresponding strengthening of users' rights is likely to lead to undermined protection, and while ensuring *ex post*

78 See, for example, Monica Horten, 'Macron-May Internet deal: Necessary measures or prior restraint?' (Iptegrity.com, 28 July 2017) http://www.iptegrity.com/index.php/internet-freedoms/1068-macron-may-internet-deal-necessary-measures-or-prior-restraint, accessed 15 January 2018.

79 Anja Mihr, Cyber Justice: Human Rights and Good Governance for the Internet (Springer, 2017), 47.

80 See, for example, Emily B. Laidlaw, Regulating Speech in Cyberspace (Cambridge University Press, 2015) Chapter 6.

facto safeguards and modalities for redress is important, it is not sufficient, particularly as existing studies indicate that these tools go underused.[81]

There are, of course, legitimate and practical justifications for stressing the role and responsibility of social media companies in the context of countering terrorism and violent extremism. Due to the control and influence they exercise over content on their platforms, meaningful action could not be taken without their cooperation. Having private actors such as social media companies increasingly undertake traditionally public tasks in the context of Internet governance is probably unavoidable, especially as public authorities (including the judiciary) in most States do not have the human or technical resources to satisfactorily perform these tasks.

While it is inevitable for relevant private actors to play an increasingly significant role, including the taking up of quasi-executive and quasi-adjudicative tasks, this should not be done without proper guidance and safeguards. At this point, however, the outsourcing results in lowering or removing existing human rights safeguards and protections. Social media companies are stuck with tasks that they are not particularly well-equipped to carry out. For example, it is questionable whether private actors are well-placed to assess whether a particular measure is necessary and proportionate in the interest of national security or public order.

Social media platforms should be given clear and detailed instructions and guidance if they are to carry out such assessments. If control over elements of the right to freedom of expression are outsourced to these outlets, independent oversight of their conduct in this respect needs to be ensured, to guarantee transparency, accountability and respect for the right to remedy of individuals whose rights are unjustly interfered with in the process. The necessity for safeguards is not simply due to intermediaries lacking the relevant legal expertise, but a basic matter of legal principle requiring that measures impacting human rights be subjected to independent oversight by public, preferably judicial, authorities rather than left up to private bodies.

The challenges that arise in this domain call for ways to bridge public and private dimensions involved in promoting and protecting human rights. This in turn would require ensuring complementarity and synergy between various systems of regulation.[82]

81 See note 44, Appendix B.
82 See note 80, 233-234.