DIGITAL FUTURES IN ASIA

Online Content Moderation: Scoping Out a Geographically Proportionate Response

HATIM HUSSAIN

IT for Change

Digital Futures In Asia Series

This essay is part of IT for Change's Digital Futures in Asia series. In a post-Covid context, the rapidly unfolding datafication of the economy has acquired heightened momentum. The Digital Futures in Asia series looks to asses and understand the impact of critical digital transformations in the Asian region in sectors such as work, finance and social media – which are at a crucial conjuncture and stand to be wholly transformed through the process of digitalization.

Published by IT for Change

November, 2022.

Title: Online Content Moderation: Scoping Out a Geographically Proportionate Response

Author: Hatim Hussain

Acknowledgements

Research Coordination: Anita Gurumurthy, Deepti Bharthur, Nandini Chami
Editorial Support: Sohel Sarkar
Proofing: Sadaf Wani
Design: Sreemoyee Mukherjee

Online Content Moderation

Scoping Out a Geographically Proportionate Response

Hatim Hussain

Contents

I. Introduction 2
1.1 Background 2
1.2. Aim of the study 2
2. The Big Challenge: Moderating Online Content 3
3. Trends and Developments in Content Restriction Practices 5
4. Extraterritorial Implications in Content Moderation 8
5. Community Guidelines and Role of Private Sector Intermediaries 12
6. Judicial Response: Takedown Orders 14
7. Developing a Framework 17
7.1. Degree of normative coherence 17
7.2. International human right standards 18
7.3. Harm 19
7.4. Audience/reach 19
7.5. Context of speech 20
7.6. Geographic relevance of the post 20
8. Legal Solutions and Recommendations 23
8.1. Convergence on the terms of service of intermediaries 23
8.2. GeoIP filtering 24
8.3. International cooperative efforts 25
Conclusion 26
Bibliography 27
Annex 1 34

I. Introduction

1.1 Background

Equal and open internet access is a widely recognized fundamental right. Yet, restrictions on content posted online and its censorship, have been on the rise, particularly on account of the Covid-19 pandemic, specifically in developing regions globally. In the third quarter of 2020 alone, Facebook censored 22.1 million posts relating to hate speech, 3.5 million on harassment, and more than 19 million posts on graphic content.(Rosen, 2020) In India specifically, more than 6,000 takedown orders were issued by the central government to social media companies.(Bhardwaj, 2021) As digital technologies become integral to our personal lives, the protection of human rights and freedom of expression on the internet has become a major transnational challenge.

Under well-established principles of jurisprudence, content restrictions must be governed by the principle of proportionality. The global nature of the internet makes online content subject to plural national laws. This jurisdictional patchwork adds to the complexity of applying the proportionality test on content restrictions, as currently there is no clarity on how to reconcile disparate domestic laws in the governance of cyberspace. Besides, the lack of an overarching global framework to reconcile national legislations adds to the complexity and renders the exercise of upholding freedom of expression in the digital space marred with controversies. It is true that content deemed legal in one country may well be declared illegal in another jurisdiction. However, restrictions to freedom of expression at the domestic level have global implications. In the absence of substantive and procedural frameworks arrived at through wide consensus, regulatory responses to digital content moderation so far have been reactive and arbitrary.

In most cases, nationwide responses to content restrictions have been diverse and in conflict with potentially varied regulatory obligations. Moreover, what kind of content is taken down is largely determined by a complex normative environment involving both state (through national laws) and non-state (through community standards) actors. This raises delicate questions concerning regulation of online content, geographic scope and reach of content restrictions, and responses that are proportionate to corresponding harm and context.

1.2. Aim of the study

Online content moderation has a strong human rights dimension. Freedom of expression and access to information are core elements of any culture – a global threat to these ideals may have a significant effect in reshaping future digital societies. This paper aims to comprehensively assess several key challenges in enforcing cross-border content restrictions. By employing a normative lens and documenting diverse technical and legal approaches, this study explores:

- How can common definitions for actions such as content blocking, takedown, stay down, withholding, or filtering be developed and implemented?

- To what extent are shared typologies useful in moderating content and how can criteria be developed that balance the twin objectives of protection of free expression and regulation of harmful/objectionable content?

- What criteria should courts use and how do we define the role of these criteria in determining the proportionate geographic scope of action?

- In what manner can responsibilities of varied stakeholders – public authorities, private intermediaries, and courts – be segregated in dealing with the large volume of harmful/ objectionable content online internationally?

This report begins by exploring the challenges to moderation of content online (section two) in four key areas: the absence of internationally-agreed-upon substantive and procedural frameworks, imprecise terminologies, difficulties in identification of illegal/ illegitimate content, and remedies. Section three charts the key overarching and topical trends in content moderation practices in India and Australia, and more generally, globally. This is done by investigating normative plurality, the expanding divide in responses to moderation by private companies and state agencies, as well as the complexities in dealing with certain forms of content such as defamation, hate speech, and extremism. At the core of content moderation lies the problem of extraterritoriality, which section four explores in greater detail. The fifth section debunks the responses at the other end of the spectrum of regulatory efforts, namely the important role of community guidelines and intermediaries in regulating content. The study then maps the geographically proportionate responses to content moderation in courts through judicial weapons such as takedown orders (section six), subsequently envisaging a regulatory framework rooted in a sixfold criteria and certain core principles which can inform courts in framing responses (section seven). The report concludes with observations on best practices and legal approaches to solutions (section eight).

2. The Big Challenge: Moderating Online Content

Private online services host billions of users who publish millions of posts and upload thousands of hours of videos every day, creating user-generated content which has become a fundamental medium for exercising the right to freedom of expression and access to information. The diversity of this content – whether it is social, political, cultural, religious, or any other – along with the sensitivities around it are increasingly maturing. Opinions that were limited to private spaces now have an unprecedented global audience and outreach, granting them not only significant visibility but also temporal permanence. While a victory for a substantive right to expression, online content can also severely harm, threaten, or offend. It is a common challenge for stakeholders responsible for protecting such rights, including states and platform intermediaries, to address the abuse caused by the presence of illegal or harmful content.

Moderating content online is undoubtedly a difficult exercise. Violations must be identified in a timely and efficient manner, while respecting human rights, and without losing sight of the goal of creating a participatory digital society that does not restrict expression. The challenge is compounded by the varied interpretations of norms governing content restrictions and associated due process safeguards on a national and regional level. In

interpreting the normative landscape, these challenges can be identified more specifically across four distinct areas:

(a) Absence of agreed upon substantive and procedural frameworks: In dealing with content moderation issues, the disparity between national legislations protecting freedom of expression and human rights online is a crucial element. As we shall see later, court decisions apply legislative provisions which often extend beyond their geographic scope, creating regulatory friction and scope for potential conflicts.(Keller, 2019) Such conflicts can further fragment regulatory and judicial responses. Instead of a single, cohesive, self-contained body of rules, we find a multiplicity of bilateral, regional, and international treaties that are subject to varying interpretations by nations at their judicial forums and at international tribunals. (Duoek, 2018)

(b) Inconsistent and vague typologies of abusive content: National legislations often employ an *a priori* scope to govern access to content, which while conforming to human rights standards, use a typology of abusive content that is inconsistent and vague on a global spectrum. Primarily, these typologies can be classified into three categories:

- Content that violates international standards and faces unanimous objection,
- Content which is objectionable but has varied interpretations on the scope of its illegality, and
- Content which is prohibited in some countries, but not in others, and even accepted in a few jurisdictions.

Employing the aforementioned criteria creates arbitrary distinctions in understanding and addressing abuses while furthering the divide in shared approaches to content regulation. This highlights the need for a global substantive harmonization and structured interactions among key stakeholders with the aim of developing an acceptable typology of objectionable content online.

Interestingly, while we observe a diversification of government obligations, there is a harmonization of global restrictions by private entities through community guidelines. Such guidelines may be deliberately envisioned to be applied globally given the jurisdiction-agnostic operations of intermediaries, but they are enforced within the context of national legislations and can diverge majorly in their effect. Within themselves, intermediary guidelines also differ due to the size, nature, and scope of their audience.

(c) Limited efficacy of mechanisms for identifying potentially infringing content: The mechanisms used to identify potentially harmful/abusive content vary internationally, with some nations employing proactive and others reactive measures. Generally, intermediaries under various national legislations are protected from liability unless they receive notice of infringing material and fail to act swiftly to remove them.[1] In such cases, platforms often deploy a range of methods and channels to detect inappropriate content, including flagging tools powered by artificial intelligence (AI) and frequent checks to enable users to submit

[1]For a detailed account of these measures, see section three of this paper.

requests to remove such content. Concerns around the effectiveness of these processes aside, it is important to note that current legislative efforts do not compel intermediaries to identify content prior to the occurrence of the infringing event.

Reactive measures in detecting online abuse raise two crucial observations. First, reliance entirely on a submission-detection-removal cycle is inadequate. It does not comprehensively deter criminals from posting infringing content online nor does it cover liability where abuse goes undetected. Second, given the high intensity of content posted online, detecting abuse proactively is often a question of scale – existing mechanisms are simply unequipped to pre-screen billions of posts on a regular basis. While use of algorithms and AI has, to a large extent, assisted in such processes, their effectiveness is still debatable. Besides, intermediaries should also be mindful of algorithmic biases and impact of false positives in automating processes.

(d) Procedural deficiencies in decision-making regarding content restrictions: Finally, procedural deficiencies are another major challenge in online content moderation. Currently, decision-making processes regarding content restrictions can be enforced either by national courts or intermediaries and, in most cases, some combination of both. For courts, specifically, adjudicating the bulk of these decisions at the first instance is not pragmatic due to limited resources. In such cases, restrictions are enforced prima facie by intermediaries with a mechanism to appeal against a decision in case of a grievance.

A crucial point of concern in existing appellate processes is the availability of redress mechanisms which are independent of reliance on intermediaries. It is not an established practice to secure appeals against decisions by intermediaries in courts. Besides, such cases of moderation have high transaction costs and there may be jurisdictional concerns when a company's decision is to be enforced at a global level.

Addressing such challenges is a collective exercise – it requires not only coordination between public and private actors, but also addressal of fundamental issues of principle. As a subsequent part of this report notes, developments in content moderation efforts in recent years have witnessed a step forward in this direction, although problems remain.

3. Trends and Developments in Content Restriction Practices

In analyzing content moderation practices, this paper observes key overarching trends in the global regulatory ecosystem, particularly in jurisdictions such as India and Australia, which lie at the center of its thematic focus. It may be prudent to analyze these trends at the macro level given the shared concerns many jurisdictions face. Nevertheless, it is important to note initially what the legal position on content moderation is in India and Australia.

In response to the livestreamed attack on a mosque in Christchurch, New Zealand in 2019, Australia promulgated an amendment to the Australian Criminal Code requiring companies to proactively moderate content by removing objectionable material "expeditiously". (Criminal Code Amendment, 2019) Large penalties were imposed on companies that fail to moderate or report such content, including jail sentences for their executives. At the policy level, Australia also imposes guidelines on businesses to moderate end users' content. Such strategies include disabling or muting account of the offenders temporarily or permanently

(depending on the severity of crime), expanding the scope of moderation to include all forms of content, enabling user participation in moderation processes (by flagging/reporting spam or inappropriate content), using automation techniques to filter keywords/phrases and equipping social media pages with moderators who regularly check accounts and/or community threads.(Walker, 2020)

In India, Section 69A of the Information Technology Act, 2000 provides legislative power to the government to issue directions to intermediaries to block access to content, based on criteria such as national sovereignty and integrity, security and defense, friendly relations with foreign states, or public order. (Information Technology Act, 2000) Notably, the Intermediary Guidelines of 2011 prescribes safe harbor protection to intermediaries from liability in select circumstances. (Intermediaries Guidelines Rules, 2011) In 2018, an amendment to these rules, similar to the Criminal Code Amendment in 2019 in Australia, were envisaged to push for proactive monitoring of content and removal of content deemed objectionable within 24 hours. (Intermediaries Guidelines Rules, 2018) The procedure to issue directions to intermediaries is provided in the Information Technology (Procedure and Safeguards for Blocking of Access of Information by Public) Rules, 2009 – one of the key features of the regulation is a restriction on public disclosure of the action taken to block content when directions are issued by the government.[2] Beyond the statutory measures, it is also important to note that the Supreme Court of India, in the landmark *Shreya Singhal v Union of India*, struck down Section 66A of the Information Technology Act, 2000, which previously criminalized the publication of offensive comments online. (Singhal v. Union of India, 2013)

> *Self-regulation functions within the realm of government responsibility and an international agreement on jurisdictional standards is crucial to allow companies to self-regulate effectively.*

While similar regulatory frameworks exist in other jurisdictions as well, what is interesting to note is the variations in the ways in which national frameworks determine the rights of netizens. In this context, one concept that deserves attention is sovereignty. Traditionally, it is much harder to prove claims of sovereignty when crimes occur in cyberspace. In *re Search Warrants Nos 16-960-M-01 and 16-1061-M to Google* (2017), a United States District Court wrestled on this issue, noting the difficulty of determining which foreign country's sovereignty is infringed upon by an act of an intermediary. (In re Search Warrants Nos 16-960-M-01 and 16-1061-M to Google, 2017) Section 4 of this report explores this perplexing question in greater detail.

For now, it is crucial to note that the regulatory environment in these jurisdictions have evolved significantly, moving beyond the question of 'regulation *of* the internet' and focusing more on 'regulation *on* the internet'. Applying of the aforementioned laws in the digital sphere is now also firmly established in courts. More importantly, courts have started undermining the centrality of regulators in governing content issues on the internet, thus thereby underscoring the importance of 'mesh regulation' or the role of varied stakeholders involved in infrastructure governance.[3] To some extent, this is inevitable given the peer-

[2] Ibid., r. 16.

[3] Some scholars have advanced the idea of a pyramidal model of regulation for the internet, with the regulator or the state at the center, but developments in technology

driven nature of content moderation and the importance of self-regulation and community guidelines imposed by intermediaries. However, as noted earlier, self-regulation functions within the realm of government responsibility and an international agreement on jurisdictional standards is crucial to allow companies to self-regulate effectively.

Beyond self-governance (which may have its own perils) (Satariano, A. & Issac, M. 2021), several other regulatory practices in content moderation have also gained prominence. These include a combination of private and public law regulation, technical codes (lex informatica) and public-private arrangements.[4] These developments are, however, 'branching in' and appear to be country-specific rather than 'branching out' to become harmonized, thereby complicating the regulatory landscape. Such regulatory initiatives by countries often range broadly within the spectrum of 'inaction' to 'overaction', (Svantesson, 2017) and can either be conflicting, coexisting, or interdependent. Response to regulation of the internet has not been static, and jurisdictions often swing on both ends (of under- and over-regulation) during different periods.

Stronger entities, such as the European Union and the United States, tend to influence decisions of states that are in the process of developing their regulatory stance, such as India, on these issues.

At present, regulatory behavior in India and Australia, as in developed jurisdictions such as the European Union and the United States, seem to be moving in the direction of over-regulation. Partly a response to the global trend, these regions are continuously seeking to expand their territoriality in cyberspace and apply their laws beyond their geographic boundaries. While ambitious, state regulation of platform environments presents enforcement challenges – complexities in legal compliance add enormous obstacles for enforcement, given the underlying confusion in applying laws on cyberspace. Besides, proper enforcement of orders of the domestic courts, which tend to venture out and apply to parties and situations beyond their ambit, is nothing more than a theoretical possibility.

There is also no normative convergence on fundamental issues such as extraterritoriality, the type of content deemed acceptable, or an internationally established taxonomy of regulatory response to illegal content. This obscurity of classificatory attempts adds fuel to the divergence of norms. As we shall see, nation-states' uncoordinated and reactive position towards these issues is a complex challenge and has detrimental impacts, which not only create friction between nations but also conflicts.

It is not surprising, therefore, that normative plurality on a jurisdictional level has helped replicate and stimulate legal approaches across these nations, resulting in a form of cross-fertilization. Stronger entities, such as the European Union and the United States, tend to influence decisions of states that are in the process of developing their regulatory stance, such as India, on these issues. An example of this is the development of case law in India, which has repeatedly cited judicial precedents from the United States to justify the extraterritorial ambit of content moderation.(Ramdev vs Facebook, 2019) Such developments are, perhaps, an outcome of the lack of regional diversity in the adjudication of complex cross-border content moderation

and globalization have severely undermined this theory. See Ost, F., & Kerchove, M.V. (2019). From the pyramid to the network: for a dialectical theory of law. Brussels: Presses de l'Université Saint-

Louis; Weitzenboeck, E. M. (2014). Hybrid net: The regulatory framework of ICANN and the DNS. International Journal of Law and Information Technology, 22(1), 49–73. https://doi.org/10.1093/ijlit/eat016

[4] Ibid.

issues, as developed regions tend to have a common denominator and thus garner greater support in discussions on such issues. Besides, it is possible to argue that since large internet companies are based in more developed countries,[5] thereby directly falling within their sphere of influence, they are more equipped to impose their laws globally. However, discussions on content moderation must be inclusive to avoid such skewed perceptions from becoming a global norm.

It is true that instances of judicial imitation are abound in other fields of law, and are usually considered 'good practice'. However, in the context of content moderation, they can lead to varied results. Adopting an aggressive approach to moderating content online (such as, for instance, imposing global takedown orders) may create more cross-border enforcement challenges than adopting an approach which reflects an increased international harmonization. Besides, there is always a threat of 'scalability' – an approach which infringes on sovereignty and self-determination of other nations is not practical. This trend of the digital sphere becoming a new battlefield for international conflict is definitely concerning.

Like regulatory behavior, the trend of normative divergence can also be observed in practice in judicial decision-making. This divergence stems from the fact that while country X may classify a category of content as prohibited, country Y may deem it acceptable. For instance, in *LICRA v. Yahoo! France* (2000), a United States Court refused to enforce a French court decision ordering injunction on operation of an auction website selling Nazi material since sale of such materials, while illegal in France, was legal in the US.[6]

From the stream of overarching and topical trends described above, one particular theme emerges consistently – extra-territoriality. The following section aims to deconstruct prevailing notions around this issue.

4. Extraterritorial Implications in Content Moderation

While the internet is a sovereign territory unrestricted by traditional jurisdictional limitations,[7] as discussed earlier, enforcement of domestic laws over cyberspace has been a subject of intense debate.(Solow-Niederman et al, 2017) Since states often misuse ambiguities around extraterritoriality to distinguish between legitimate and illegitimate content moderation cases,[8] it is crucial to explore the threat of the

[5] This is particularly prevalent for companies based in the United States, including Facebook, Amazon, Google, Apple, and Twitter. Other non-western instances of this can be found in China which has exported its Chinese norms while providing subsidized mobile and broadband services in Africa.

[6] Yahoo!, Inc. v. La Ligue Contre Le Racisme et L'Antisemitisme, 145 F. Supp. 2d 1168,1171 (N.D. Cal. 2001). http://www.lapres.net/ya2011.html. See also, Greenberg, M. (2003). A return to lilliput: the LICRA v. Yahoo! case and the regulation of online content in the world market. Berkeley Tech Law Journal. 18 Berkeley Tech. L. J. 1191 (2003). https://digitalcommons.law.ggu.edu/cgi/viewcontent.cgi?article=1430&context=pubs

[7] Ryngaert, C. (2015). Jurisdiction in International Law. 2nd edn. Oxford: Oxford University Press. https://opil.ouplaw.com/view/10.1093/law/9780199688517.001.0001/law-9780199688517. See also, Schmitt, M. (Ed.). (2017). Tallinn manual 2.0 on the onternational law applicable to cyber operations. Cambridge: Cambridge University Press. https://assets.cambridge.org/97811071/77222/frontmatter/9781107177222_frontmatter.pdf (p. 61-72).

[8] As per principles of international law, a clear consensus on extraterritorial jurisdictional claims is absent. In Microsoft Warrant, the court noted that a concise definition of extraterritoriality is non-existent even in legal systems which prescribe an express presumption against illegality. This is further evidence of the already increasing dichotomy in using territoriality as a principle to deal with cross-border legal challenges in regulating content moderation. See, Nojeim, G. (2014, July 30). Microsoft Ireland case: Can a United States warrant compel a United States provider to disclose data stored abroad?. Centre for Democracy and Technology. https://cdt.org/insights/microsoft-ireland-case-can-a-us-warrant-compel-a-us-provider-to-disclose-data-stored-abroad/

extraterritorial dichotomy in cross-border legal challenges to content restrictions.

The historical adherence to territoriality as a foundation of jurisdiction under legal systems has witnessed a slow death over the years.[9] In itself, the concept of territoriality has multiple limitations when applied to digital spaces. Its two core functions of (a) providing states a criterion to claim jurisdiction and (b) limiting the exclusive domain of nation states (while protecting sovereignty of others), are inoperable in cyberspace. This is because any state can claim jurisdiction online by finding a territorial anchor point,[10] but it is simply unrealistic to expect sovereign exclusivity in a digital world where states are part of a global community. Under such circumstances, extending the principle of territoriality to the regulation of content restrictions may lead to serious drawbacks.

While territoriality is no longer relevant, courts have sought to reimagine the application of laws beyond national boundaries by extending the regulatory ambit to individuals and organizations in other states in order to address transnational issues that have some nexus with national laws. However, competing interests of nation states in pursuing a content takedown request beyond the contours of their own territory pose at least two distinct risks: on the one hand, extraterritorial enforcement directly jeopardizes the freedom of expression of citizens of a nation state at the recipient end of the sanction request, and on the other, it contradicts established rules of comity under international law. In such cases, resolution of content moderation issues by courts leads to an unsavory choice which favors national sovereignty over freedom of expression.

> *Disproportionate responses to content moderation are an outcome of deficiencies in the enforcement of national court decisions.*

The prevalent conflict between domestic laws on the internet and their international application raises another crucial challenge. Since laws applicable to content available on the internet are enforced on a national basis, their identification and interpretation is left to intermediaries, which results in major challenges. A global consensus over what form of content is unacceptable is absent, and national legislations significantly differ on their classification of legal and illegal text (see section six of this paper). Moreover, the extent of criminal penalties on such offences (viz. hate speech, blasphemy, harassment, defamation, etc.) also vary enormously. In such cases, concerns regarding public order can often lead to an abusive invocation of excessive removal requests, including internet shutdowns and blocking of internet service protocols (ISPs). Precedents of such disproportionate blocking can be seen in

[9] Examples of this decline can be found in, for instance, the proposal of the European Union for a directive which seeks to establish harmonized rules for gathering evidence in cross-border criminal proceedings; the 2018 United States CLOUD Act; EU's proposal for regulation of electronic evidence; 2nd Addl Protocol Amending Budapest Convention and the GDPR's Article 3(1) (which lays down that location for the purposes of data processing is not relevant). These instruments broadly denote a shift from the principle of territoriality.

[10] In the opinion of the Advocate General in Concurrence Sàrl v Samsung Electronics France SAS and Amazon Services Europe Sàrl, it was noted that the internet, by definition, is universal and the causal event or damages sustained are very difficult to determine. *See,* Opinion of Advocate General Wathelet in General in *Concurrence Sàrl v Samsung Electronics France SAS and Amazon Services Europe Sàrl*. ECLI:EU:C:2016:843 (2016). https://curia.europa.eu/juris/liste.jsf?language=en&num=C-618/15

India and elsewhere.[11]

Beyond an uncertain application of national legislations internationally, extraterritoriality is also contentious from an enforcement perspective. Disproportionate responses to content moderation are an outcome of deficiencies in the enforcement of national court decisions. To a large extent, courts in different jurisdictions adopt nationally-determined criteria to regulate content beyond their jurisdictional ambit. This has triggered inconsistent reactions, including, at times, circumstances where decisions are contrary to those in other jurisdictions. For instance, in 2017, the Mexican Supreme Court invoked jurisdiction over Google solely on the ground that the actions of the company impacted the fundamental rights of Mexican citizens.(Riquelme & Galindo, 2017) Google's argument that its operation was based in the United States and that it had no physical operations in Mexico was refused by the courts as its actions, while remote, still infringed on human rights.

Remarkably, such jurisdiction issues are a huge subset of a range of content restriction cases adjudicated in courts. In the context of hate speech, these issues arise due to the lack of an acceptable position on distinctions between defamatory content and legitimate political content, internationally.[12] Cross-border defamation is another category which raises extraterritoriality concerns due to the widely recognized struggle between the exercise of freedom of expression and the restrictions envisaged to protect the right of reputation of others.[13] In such cases, a question often arises regarding the geographic scope of damages, which may either be national, regional, or global in nature.(Cases C-509/09 and C-161/10, 2011) Courts have, indeed, found it troublesome to examine cases where awarding damages beyond their jurisdiction potentially interferes with the right to freedom of expression in other states. For instance, in *Dow Jones v Gutnick* (2002), a landmark case in Australia, the High Court limited the geographic scope of damages sought by the plaintiffs to avoid interpreting liability of damages occurring in other nations. In para 54, the court observed:

> "…the spectre which Dow Jones sought to conjure up in the present appeal, of a publisher forced to consider every article it publishes on the World Wide Web against the defamation laws of every country from Afghanistan to Zimbabwe is seen to be unreal when it is recalled that in all except the most unusual of cases, identifying the person about whom material is to be published will readily identify the defamation law to which that person may resort."(Dow Jones and Company Inc v Gutnick, 2002)

[11]See for instance, Press. (2021, April 26). India Covid: Anger as Twitter ordered to remove critical virus posts. *BBC.* https://www.bbc.com/news/world-asia-56883483. For a more detailed overview of the cases of internet shutdowns globally, see also, Taye, B. (2021) Shattered dreams and lost opportunities. *Access Now*. https://www.accessnow.org/cms/assets/uploads/2021/03/KeepItOn-report-on-the-2020-data_Mar-2021_3.pdf

[12]For instance, hate speech is accorded a higher degree of protection in the United States than in regions such as Germany, Canada, and Europe. *See*, Brugger, W. (2002). Ban On or Protection of Hate Speech? Some Observations Based on German and American Law. *Tulane European and Civil Law Forum.* https://journals.tulane.edu/teclf/article/view/1662/1494

[13] *See*, International Covenant of Civil and Political Rights. art. 19(3). https://treaties.un.org/doc/publication/unts/volume%20999/volume-999-i-14668-english.pdf. In June 2019, this issue was also raised in Australia's Council of Attorney's General Review of Model Defamation Provisions. *See,* Model Defamation Provisions (2020). *Parliamentary Counsel's Committee.* https://pcc.gov.au/uniform/2020/Consolidated_Model_Defamation_Provisions.pdf

More generally, in the context of the internet, and specifically in cases where courts are obliged to respond to requests of worldwide deletion or rectification of defamatory content, questions on the scope of jurisdiction are amplified.[14] Akin to developing regions, evolved jurisdictions such as the European Union have also found it perplexing to regulate the removal of information globally. (Glawischnig-Piesczek v. Facebook Ireland Limited, 2019) In *Glawischnig-Piesczek* (2018), the court was quick to note that while member states in the European Union may adjudicate removal of information worldwide, the courts must be cautious to recognize the differences in national laws and adopt "an approach of self-limitation".[15] Notably, the Court of Justice of the European Union (CJEU) also recommended that courts in member states limit extraterritorial effects to avoid any damage to personality rights and private lives of persons, while also observing adherence to principles of international comity and doing only what is "strictly necessary".[16] CJEU's decision in *Glawischnig-Piesczek* is crucial in recognizing that attempts at extraterritorial enforcement are seen as options of last resort.

While the European Union follows a restrictive approach to extraterritorial enforcement, Australia has done the opposite. A high-profile case of sexual assault by a cardinal of two choirboys raised questions on jurisdictional aspects in cases of press freedom and suppression of content globally.(Paul, 2021) International news outlets were penalized for publishing details on sex abuse.[17] In India, courts have used similar reasoning to justify removal of content, albeit globally, with wide extraterritorial ambit. In 2019, the High Court of Delhi ordered three major platforms, Twitter, Google, and Facebook, to remove defamatory content globally citing violation of domestic laws.[18]

An overly ambitious extraterritorial application of rules by courts risks leaving the aggrieved without judicial redress facilities and imposes significant cost of compliance. It is also impractical to expect an internet user to know the laws of all nations to be able to comply with them. Besides, under established rules of international law, a state is bound by some restrictions in claiming enforcement of its laws. There are also other concerns surrounding arbitrariness, unrequited contingencies, and uncertainty in extending juridical ambit extraterritorially that courts must be mindful of.

Despite these problems, the courts have been entrepreneurial in exploring solutions to address the complexities that arise in the exercise of personal jurisdiction in other nations. As early as in 1997, a United States court devised a 'sliding scale' test to examine extraterritorial ambit by noting that personal jurisdiction of the court is directly proportional to the internet website's commercial activity, that is, the more active the website is, the more legitimate an extraterritorial claim will be.(Zippo Manufacturing Co. v. Zippo Dot Com, Inc, 1997) On the other

[14] See, Bolagsupplysningen OÜ and Ingrid Ilsjan v Svensk Handel AB. ECLI:EU:C:2017:766 (2017). https://curia.europa.eu/juris/liste.jsf?num=C-194/16 (para 50). See also, Calster. V. G. (2017). Close, but no sigar. The CJEU on libel, internet and centre of interests in Bolagsupplysningen. https://gavclaw.com/2017/11/15/close-but-no-sigar-the-cjeu-on-libel-internet-and-centre-of-interests-in-bolagsupplysningen/

[15] *See,* Opinion of Advocate General Szpunar in ibid 31. https://curia.europa.eu/juris/document/document.jsf?text=&docid=214686&pageIndex=0&doclang=EN&mode=req&dir=&occ=first&part=1&cid=7255939 (para 100). *See also,* Keller, D. (2019). Dolphins in the Net: Internet Content Filters and the Advocate General's Glawischnig-Piesczek v. Facebook Ireland Opinion. *Stanford Centre for Internet and Society.* https://cyberlaw.stanford.edu/files/Dolphins-in-the-Net-AG-Analysis.pdf

[16]Notably, use of technical measures such as geoblocking to limit the effect of the information have also been recommended by the court. See Section 8 of this report for a detailed analysis.

[17] Ibid.

[18]Supra n 16

hand, Australian courts have sought the solution to jurisdiction issues in the doctrine of *forum non conveniens* – the principle that jurisdiction may not be exercised by courts where adjudication in another court is more convenient.[19] Notably, a two-factor standard, which relies on (a) legitimate interest and (b) substantial connection to subject matter of dispute (that is, strong nexus), along with an equivalent assessment of consideration of other interests might prove to be an effective tool to moderate content.

Nevertheless, while such solutions may be unique, there is a need for stakeholders to collaborate on cross-border moderation issues online. Subsequent parts of this report focus on evolving a framework to address such challenges, as well as the role of community guidelines established by private sector intermediaries in resolving disputes on content restrictions.

5. Community Guidelines and Role of Private Sector Intermediaries

Liability on private operators as 'gatekeepers' of content has risen steadily as internet platforms become primary content providers online. As guardians of freedom of expression, their responsibility is particularly notable in the context of hate speech, extremism, and terrorism which require an active response to content moderation.[20] Intermediaries assume a range of functions on all three levels: acting as lawmakers defining what constitutes legal and illegal content, as administrators blocking content deemed illegal, and as judges who adjudicate on the legitimacy of content. Since decisions relating to moderation of content online are enforced *prima facie* at the platform level, given the breadth and intensity of impact it can have on free speech, it is helpful to scope current practices on content moderation in the private sector.

Content sharing platforms thrive in enabling a democratic forum which allows users to share, create, distribute, and critique content publicly. As facilitators of free speech, their cross-border implications are obvious.(Elkin-Koren & Perel, 2020) However, given their dual role as private, profit-maximizing entities driven by financial incentives who also govern speech, balancing commercial interests with adjudication of content that advances public welfare and protects public interests poses an interesting challenge.

Regardless of the size, most online platforms regulate content through terms of service or community guidelines which are jurisdiction-agnostic rules to moderate users' behavior.(De Street, A. et al. 2020) However, given that they operate within the contours of national legislations, they vary significantly from one jurisdiction to another, often regulating or adjudicating on the same content differently. Specific measures deployed by intermediaries to identify illegal content include flagging such content ('notice and takedown' schemes), using algorithmic decision-making with

[19] Supra n 29. *See also,* Haaretz.com v. Goldhar. 2018 SCC 28. https://scc-csc.lexum.com/scc-csc/scc-csc/en/item/17115/index.do

Supra n 32 at para 100 (on how courts must adopt an approach of 'self-limitation').

[20] Research Report by the Mandate of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, with the support of the International Justice Clinic at the University of California, Irvine School of Law. (2020, July 30). *Office of High Commissioner for Human Rights.* https://www.ohchr.org/Documents/Issues/Opinion/ResearchPaper2020.pdf

human intervention to moderate content, or deploying keyword filters that restrict use of objectionable language. Legal systems in both India and Australia require platforms to identify and remove illegal content at varying speeds for different types of content, including within a span of 24 hours in certain cases.(Criminal Code Ammendment, 2019)

Of particular significance is the role of intermediaries in performing quasi-judicial functions, the majority of which have important human rights implications. Assuming additional responsibilities to systematically monitor and remove content, thus, raises crucial questions regarding the criteria and procedures employed, while giving intermediaries unsolicited power to enforce community guidelines in accordance with their best interests. Examples of such capacity are found in legal systems in Australia,[21] France (Law no. 2020-766, 2020), Brazil,[22] Bangladesh (Digital Security Act, 2018), Tanzania (Electronic and Postal Communications Regulations, 2020), Uganda (Namubiru, 2018), Singapore (Protection from Online Falsehoods and Manipulation Act, 2019) and many other nations. As we discuss later, current legislative frameworks do not provide adequate clarity or structure to propel platforms to adopt an overly inclusive strategy to restrict content or limit access in a wider geography to be 'safe' from violation. (Kaye et al, 2019) Vagueness in definitions of illegal content such as hate speech or extremism only complicate enforcement and force intermediaries to err on the side of caution.

> *No overarching international principle currently provides clarity on which laws to abide by, and intermediaries may find it convenient to comply with duties in one state and restrict rights in another to avoid liability. This leads to a race to the bottom and undermines free and fair access to the internet*

In focusing on content moderation practices in relation to internet intermediaries, it is crucial to underscore the fact that platforms potentially regulate large quantities of content which impact fundamental rights of users. Given the cross-border effect, legitimate issues also arise in the process of enforcing legal rules, which must be addressed. Foremost among them is a situation in which an intermediary risks violating legal rules of one jurisdiction to comply with the rules of another nation. For instance, while an intermediary in Germany under *German Network Enforcement Act (NetzDG)* (Network Enforcement Act, 2017) is required to limit hate speech only in regions where the risk of proliferation is high, the French *Avia Law* (Schuler & Znaty 2020) allows an expansive interpretation of hate speech. Such country-specific requirements complicate enforcement of content, resulting in a conflict of laws. Implementing localized solutions also risks fragmentation of the internet, with a

[21] Supra n 42. s. 474.33. See also, Mandates of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression; and the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism. (2019, April 4). *Office of High Commissioner for Human Rights.* https://spcommreports.ohchr.org/TMResultsBase/DownLoadPublicCommunicationFile?gId=24533

[22] Bill No. 2,630/2020 on "Freedom, responsibility and transparency on the internet (Br.). https://legis.senado.leg.br/sdleg-getter/documento?dm=8127649&ts=1593563111041&disposition=inline. *See also*, Alimonti, V. (2020, 7 June). New hasty attempt to tackle fake news in Brazil heavily strikes privacy and free expression. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2020/06/new-hasty-attempt-tackle-fake-news-brazil-heavily-strikes-privacy-and-free; GNI expresses concern about proposed 'fake news' law in Brazil. *Global Network Initiative.* https://globalnetworkinitiative.org/gni-concerns-brazil-fake-news-law/

diverse set of rules applicable in different regions.[23]

On other fronts, it is necessary to clarify a framework which allows intermediaries to determine the appropriate geographic scope of jurisdiction in moderating content (see section seven of this report). There also seems to be a need to distinguish between an intermediary's role as a 'publisher of content' and as a 'neutral platform adjudicating on content' since protections for neutral platforms may not transfer to intermediaries acting as publishers. Doing so would help reduce bias (such as platforms promoting specific narratives for political gains), and guide platforms on what is expected of them in such situations.

Addressing the challenges enumerated above is difficult, and requires a restructuring of the fundamental principles that govern our domestic laws. For instance, rights accorded to an individual in a nation may not correspond to a legal duty in another, creating a conflict in determining which nation's law holds 'priority' over the other. No overarching international principle currently provides clarity on which laws to abide by, and intermediaries may find it convenient to comply with duties in one state and restrict rights in another to avoid liability. This leads to a race to the bottom and undermines free and fair access to the internet.[24] Beyond clarity in international frameworks, there are also best practices in certain jurisdictions which can be emulated. For instance, Section 6A of the Australian Privacy Act[25] places a limitation on the extraterritoriality of the law by recognizing that an act does not breach privacy if it is "required by the applicable law of the foreign country" (see section eight for a detailed discussion on this topic).

6. Judicial Response: Takedown Orders

So far, this report has attempted to deconstruct normative practices on a macro level, including those prevalent in select jurisdictions such as Australia and India, to moderate content online. As highlighted earlier, disparities in national legislations and the absence of globally-agreed-upon substantive and procedural frameworks that uphold freedom of expression and human rights, create ripple effects. Before charting the normative solutions to this crucial debate, it is worthwhile to examine how content moderation orders are enforced in courts so as to canvas and comprehend the geographic scope of moderation practices and envisage policy solutions.

Globally, content takedown orders are the most prevalent mode of restrictions and cover a range of content, including hate speech, privacy, distribution of objectionable content, misinformation, bullying, pornography, defamation, and fraud. (Svantesson,

[23]Supra n 41 at 61.

[24]Supra n 41 at 53

[25] Privacy Act, 1988. s. 6A. http://www5.austlii.edu.au/au/legis/cth/consol_act/pa1988108/s6a.html#:~:text=(1)%20For%20the%20purposes%20of,or%20inconsistent%20with%2C%20that%20principle.&text=(b)%20the%20act%20or%20practice,is%20inconsistent%20with%20the%20principle (states "6A (4) An act or practice does not breach an Australian Privacy Principle if: (a) the act is done, or the practice is engaged in, outside Australia and the external Territories; and (b) the act or practice is required by an applicable law of a foreign country".

2019) Such orders have also been enforced to suppress political and religious speech and impose restrictions on freedom.(Volokh, E. 2015) Broadly, a takedown order requires content declared illegitimate to be removed from the internet either locally, regionally, or globally. Another category of orders called a 'stay-up' order provides for reinstatement of content that has been blocked, delisted, deleted, dereferenced, deindexed, removed, or taken down.(Keller, 2019) Conversely, a 'stay-down' order deals with illegitimate content by preventing its publication instead of taking it down, thus placing an obligation on a publisher to proscribe certain types of content from being uploaded online.[26] Stay down orders are rare given the practical concerns in their implementation, majorly arising due to the restrained capacity of the platforms to filter *every* content published on the internet.

Section four of this report highlighted how judicial restraint has been on a declining trend in recent years, as courts engage in a race to extend application of laws beyond their geographic boundaries. In the context of takedown orders, a high-profile instance of this is the landmark 2017 *Equestek* case (Google Inc. v. Equustek Solutions Inc, 2017), where the Supreme Court of Canada ordered Google to remove the aggrieved Canadian company's websites from its servers globally. Numerous other takedown orders have since been enforced around the world,[27] which reflect *Equestek's* reasoning. In the European Union, the CJEU in *Glawischnig-Piesczek*[28] declared that the European Union E-Commerce Directive (Directive on Electronic Commerce, 2000) did not preclude member nations from ordering removal of content worldwide since Article 18(1) of that Directive did not place a "territorial limitation" [unlike Article 6(1) of the Australian Privacy Act] to prohibit extraterritorial impact of the legislation.

> *Even when a court adjudicates based on the territoriality principle, the effect of a decision may nevertheless turn out to be extraterritorial. Such consequences, while unintended or unintentional, reflect how technological realities render traditional legal principles obsolete and raise implementation concerns.*

Accordingly, determination of the geographic scope of the restriction was left to nations, with the caveat that any such restriction fell within the framework of international law (as Section 3 notes, consensus on a framework is simply absent). Another CJEU case, *Google v CNIL* (Google v. CNIL, 2019), while adjudicating on the application of territoriality in 'right to be forgotten' cases recognized that an enforcement of a dereferencing order can only be done substantively when its territorial impact is global.[29] However, given the absence of a mechanism to resolve the issue of determining scope of dereferencing outside the European Union, it restricted the application on a regional level to European Union member states only.

[26] Enforced by the Court in *Glawischnig-Piesczek.* Supra n. 31.

[27] *See,* Hassel v. Bird. 420 P.3d 776. https://law.justia.com/cases/california/supreme-court/2018/s235968.html; Google Spain SL v. Agencia Española de Protección de Datos (AEPD). Case C-131/12 ECLI:EU:C:2014:317. https://tinyurl.com/vkf575uh; Russell, J. (2015, April 10). Japanese Court Orders Google To Delete Critical Reviews From Google Maps. *TechCrunch.* https://techcrunch.com/2015/04/10/japan-google-maps-reviews/; Koslov, V. (2015, August 10). Warner Bros. Wins Case Against Russian Internet Pirates in Landmark Ruling. https://www.hollywoodreporter.com/business/business-news/warner-bros-wins-case-russian-814421/

[28] Supra n 31.

[29] Ibid. at para 58.

It is not surprising to observe precedents set in developed jurisdictions influencing court orders in Asia as well. In India, while *Shreya Singhal v UOI*[30] in 2015 cautioned that application of extraterritoriality should be narrowly construed such that national standards should not be 'imposed' internationally, Indian courts have favored global takedown orders to fully enforce legal provisions.[31] For instance, in *Ramdev v. Facebook*, Delhi High Court cited *Equustek* and *CNIL* to issue an injunction against Facebook, directing the intermediary to take down and block from its platform, worldwide, alleged defamatory content posted on its website[32]; in *YouTube v. Geeta Shroff* (YouTube, LLC v. Geeta Shroff. 2018), the High Court ordered Google to take down offending materials from its website globally. Similarly in Australia, higher courts have repeatedly ordered worldwide injunctions proscribing content.[33]
What is also surprising is the Australian court's interpretation of the application of extraterritoriality principle in *X v Twitter*, where the court regarded such authority as "discretionary" and open for courts to interpret in a manner which best remedies victims.[34] The Supreme Court of New South Wales also issued a stay-down order against Twitter to oblige the platform to apply filtering and ensure the disputed information was not posted, or in case it was posted, it was removed immediately.[35]

Even when a court adjudicates based on the territoriality principle, the effect of a decision may nevertheless turn out to be extraterritorial. Such consequences, while unintended or unintentional, reflect how technological realities render traditional legal principles obsolete and raise implementation concerns. Such a scenario may potentially arise when a court's order to seize assets of a platform/intermediary in its own jurisdiction may result in its global inoperability, thus affecting rights of citizens in other jurisdictions.(Schechner et al, 2012) In such a scenario, it is likely that global content restrictions might result in countries adopting the most restrictive approaches, particularly in grey areas such as hate speech.[36]

Nevertheless, while it is appropriate to view global content restrictions as legitimate where they relate to content that is considered universally illegal (such as child sexual abuse/pornography), courts seem to disregard the fact that forms of content which have global harmonization on their legal treatment are generally rare – other forms of content are simply subjected to a range of differing norms and principles in national contexts and are, therefore, much harder to adjudicate upon. On balance, disproportionate or excessive takedown requests are, thus, more prevalent in cases where there is substantive consensus on the illegal nature of content, such as copyright infringement and piracy cases.[37] As observed in the next section , a

[30] Supra n 11.

[31] Ibid.

[32] Supra n 16.

[33] *See* X v Y & Z [2017] NSWSC 1214; X v Twitter Inc [2017] NSWSC 1300 http://www8.austlii.edu.au/cgi-bin/viewdoc/au/cases/nsw/NSWSC/2017/1300.html (para 36).

[34] Ibid. at para 36.

[35] Ibid. *See also*, Pyburne, P., & Jolly, R. (2014). Australian Governments and dilemmas in filtering the Internet: juggling freedoms against potential for harm. *Parliament of Australia.* https://parlinfo.aph.gov.au/parlInfo/download/library/prspub/3324230/upload_binary/3324230.pdf;fileType=application/pdf

[36] Supra n. 56 at p. 154.

[37] *See,* Akdeniz v. Turkey (dec.). No. 20877/10 (2014). https://hudoc.echr.coe.int/app/conversion/pdf/?library=ECHR&id=002-9493&filename=002-9493.

relatively low degree of harmonization internationally on typologies of content often propels differing treatment by various courts.

While takedown orders are powerful legal weapons to moderate content online, it is essential to have a consensual framework which regulates its extraterritorial effects. The geographic scope of such restrictions is context-specific and best determined according to the typology it fits; a blanket policy does no good to prevent conflict of laws. The subsequent section focuses on establishing a framework which strengthens a trend towards harmonization and effective enforcement of judicial decisions.

7. Developing a Framework

We can conclude by now that policy options to deal with content moderation cannot exist in isolation. Substantive norms must be applied in consonance with agreed upon international human rights, and work alongside a diverse set of national laws and private sector implementations of community guidelines/procedures. While the previous parts of this paper have delved into each of these aspects, determining coexistence and overcoming potential conflicts remains a fundamental challenge. One also cannot undermine the importance of establishing due process standards to be applied in cross-border conflicts, and developing proper appellate mechanisms that allow individuals to escalate content moderation decisions by intermediaries to judicial forums.

In order to establish consistency and effectiveness in extraterritorial content restrictions, a framework which makes a high-level assessment about content based on the following six-part threshold test can prove helpful.

7.1. Degree of normative coherence

Ascertaining convergence among legislations internationally on content that is deemed illegal is crucial for determining the geographical scope of moderation practices. Courts can assess this on a case-by-case basis, keeping in mind normative coherence globally on the legality of content, which could follow a classification system on the typology of content agreed upon internationally.

Accounting for international normative consistency, a piece of content deemed

---

pdf&TID=thkbhnilzk; UPC Telekabel Wien GmbH v Constantin Film Verleih GmbH and Wega Filmproduktionsgesellschaft mbH. C-314/12 (2014). https://curia.europa.eu/juris/liste.jsf?num=C-314/12; Telcinco v. YouTube (2014). https://www.poderjudicial.es/search/AN/openCDocument/c5308f76d64f5ce5764284b82ec53024179e3f439af7b2cc; GEMA v Deutsche Telekom. I ZR 174/14. (2014). http://juris.bundesgerichtshof.de/cgi-bin/rechtsprechung/document.py?Gericht=bgh&Art=en&nr=73491&pos=0&anz=1; Anonymous plaintiff v. Higher Regional Court Celle (Oberlandesgericht Celle). ECLI:DE:BVerfG:2019 (2019). https://fra.europa.eu/en/caselaw-reference/germany-federal-constitutional-court-1-bvr-27617-right-be-forgotten-ii; Key Systems Case (Germany). (2014). https://www.raschlegal.de/uploads/media/LG_Saarbruecken__Urt._v._15.01.2014__Az._7_O_82-13.pdf; Stichting Brein v. Ziggo BV. C-61-/15 (2015). https://curia.europa.eu/juris/document/document.jsf?text=&docid=191707&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=2064390; Dramatico Entertainment Limited v British Sky Broadcasting Limited. [2012] EWHC 268 (Ch). https://www.bailii.org/ew/cases/EWHC/Ch/2012/268.html; Twentieth Century Fox Film Corporation And Others V Sky Uk Ltd And Others. [2015] EWHC 1082 (Ch). http://www.bailii.org/ew/cases/EWHC/Ch/2015/1082.html; Asociación de Gestión de Derechos Intelectuales (AGEDI) vs Neij Holdings LMT-The Pirate Bay. (2015). https://www.ibtimes.com.au/pirate-bay-blockade-spain-spanish-court-orders-isps-block-tpb-1434455 and Goear (2015). http://laadministracionaldia.inap.es/noticia.asp?id=1138997

illegal under a local law can thus be adjudicated on a *prima facie* level by intermediaries themselves,. In general, responses to illegitimate content are limited to the local level (through use of techniques such as geoIP) in cases where there is a lower degree of normative coherence. In exceptional cases, such as child sexual abuse or copyright infringement, where there exists a high convergence on global norms, a global restriction which meets a higher threshold can be implemented. Similarly, when a content is in violation of community guidelines, geographic responses can be directly proportionate to the degree of convergence on international norms.

7.2. International human right standards

Courts must also keep in mind the principles of international human rights as applicable to content restrictions which have a global/regional scope, and make determinations that advance the principle of international comity.[38] Adherence to international law achieves twin purposes: protecting against judicial excesses and attaining a degree of harmonization in content restriction norms. Under existing human rights law, Article 19(1) of the International Covenant on Civil and Political Rights (ICCPR) protects the right of a citizen to hold opinions, while freedom of expression is guaranteed under Article 19(2) regardless of the medium of expression or ideas.(ICCPR, 1986) Various other global[39] and regional mechanisms (Handyside v. the United Kingdom, 1986) exist which reinforce this right.[40]

International human rights standards provide a framework for governments and intermediaries alike to approach violations in online expression. As a first principle, nations should not place restrictions on intermediaries to limit expression that are precluded under international human rights law. For instance, Article 19(3) of ICCPR proscribes certain forms of content and actions, such as criminalization of political dissent and internet shutdowns, yet we see such practices unfold in various jurisdictions, particularly in India.[41] Second, clarity and precision in identifying content deemed unlawful as required under Article 19(3) of ICCPR (ICCPR, 1986) would help constrain excessive judicial intervention particularly in cases where the risk of improper identification is high.[42] Finally, regulation of online intermediaries should comply with

[38] For instance, Indian Constitution (Directive Principles of State Policy) advances principle of international comity. *See,* Constitution of India, 1951. art. 51. https://indiankanoon.org/doc/854952/

[39] *See,* for instance, Convention on the Rights of Persons with Disabilities. art. 21. https://www.un.org/disabilities/documents/convention/convoptprot-e.pdf; European Convention on Human Rights. art. 10. https://www.echr.coe.int/documents/convention_eng.pdf; Convention on the Rights of the Child. art. 13. https://www.ohchr.org/documents/professionalinterest/crc.pdf; International Convention on the Elimination of All Forms of Racial Discrimination. art. 5. https://treaties.un.org/doc/Treaties/1969/03/19690312%2008-49%20AM/Ch_IV_2p.pdf; the International Convention on the Protection of the Rights of All Migrant Workers and Members of Their Families. art. 13. https://www.ohchr.org/Documents/ProfessionalInterest/cmw.pdf; American Convention on Human Rights. art. 13. https://www.oas.org/dil/treaties_b-32_american_convention_on_human_rights.pdf; African Charter on Human and People's Rights. art. 9. https://treaties.un.org/doc/Publication/UNTS/Volume%201520/volume-1520-I-26363-English.pdf;

[40] A model on digital policing which is grounded in human rights language is the EU's Digital Services Act, currently proposed. See, Proposal for a Regulation of the European Parliament and of the Council on contestable and fair markets in the digital sector (Digital Markets Act). European Parliament. (2020). https://ec.europa.eu/info/sites/default/files/proposal-regulation-single-market-digital-services-digital-services-act_en.pdf

[41] Supra n 24. See also, Press Release. (2019). UN rights experts urge India to end communications shutdown in Kashmir. *United Nations Office of the High Commissioner for Human Rights.* https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=24909

[42] See, e.g., Impact of measures to address terrorism and violent extremism on civic space and the rights of civil society actors and human rights defenders - Report of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while

the norms of legitimacy,[43] non-discrimination, and proportionality[44] as contained in human rights law.

7.3. Harm

Courts must also strive to shift from the currently prevailing subjective criteria for content moderation to use of a relatively more objective criterion such as 'harm'. This may entail discussing the extent to which the audience of the content is aggrieved, assessing whether the audience had the means to act on the incitement caused by the content, the damage it caused, and if the remedy prescribed addresses the harm caused.

Relying on the notion of harm as a factor in court assessments shall facilitate a proportionate response in line with human rights standards prescribed above. 'Harm', in this objective context, might not only mean harm which is caused *directly and tangibly* but also that which is *indirect or consequential*. Such a harm principle would place the ability to intervene only on actions that pose a significant risk of harm to individuals within the court's jurisdiction, thereby evading the need to analyze the 'intent' of the offence. This is particularly due to the fact that harm is objectively verifiable, unlike intent, which places a higher evidentiary threshold on courts.[45]

While the notion of 'harm' is unlikely to be entirely objective (especially when courts tend to conflate each other), it is still fraught with relatively fewer subjectivities. Some of the concerns posed by the 'harm' criterion include the difficulties involved in defining the term as well as its varied interpretations, which inject it with a certain degree of subjectivity. While each of these concerns seemingly echo worries associated with the determinative modes of assessment currently used by courts, it is hoped that the communal understanding produced with reference to, and relying upon, a shared framework envisaged in this report shall be critical in bridging the chasms.

7.4. Audience/reach

This constitutes elements such as the extent or reach of the content, its magnitude, public nature/form, as well as its audience. Courts must also consider the means of dissemination used (such as via internet or broadcast) and its frequency, besides other aspects such as the extent and quantity of dissemination and whether the circulation of content is available widely to the public or restricted to a certain audience.

7.5. Context of speech

Context holds great importance in assessing whether a content is likely to be deemed illegal or discriminatory. It has a direct impact on both causation and intent, both of

countering terrorism. (2019). A/HRC/40/52. *United Nations General Assembly.* https://documents-dds-ny.un.org/doc/UNDOC/GEN/G19/057/59/PDF/G1905759.pdf?OpenElement (para 75). This report states that often hate speech and extremism are used interchangeably even though the term in not legally defined.

[43] For instance, content that is not subjected to prohibition under Article 19 or 20(2) of the ICCPR should not be restricted (such as overly broad interpretations of religious speech or content that instigates hatred against the regime).

[44] See Section 7.A. of the report.

[45] The notion of harm as a liberty enhancing principle was advanced by John Stuart Mill in 1960s. *See,* Mill, J.S. (1963). *Collected Works of John Stuart Mill, ed. John M. Robson,* 323 vols. Toronto: University of Toronto Press, London: Routledge and Kegan Paul.

which are difficult to determine objectively. Consequently, the analysis of content in courts and the decision on the extraterritorial reach of damages should place any speech within the context of its social or political situation, prevalent at the time the speech was made. While determining context is a subjective assessment, with potentially multiple interpretations, a combination of the elements enumerated above will guide courts in reaching a normatively coherent outcome.

7.6. Geographic relevance of the post

Analysis by courts may also incorporate the degree to which the content is provocative for the audience it is intended to address. As stated earlier, content deemed illegal in one territory may be permissible in another. In such cases, moderation of content by courts must respond in a restricted context and declare penalties limited to the geographic relevance of the content.

To give effect to the parameters mentioned above, agreement on certain key principles in content moderation is essential. The following guiding principles will help determine the extra-jurisdictional extent of takedown orders.

A. Proportionality in content restrictions

While cases dealing with restrictions on freedom of expression are generally determined on principles of necessity and proportionality, given the global accessibility of online content and the application of multiple laws with varied regulatory responses and obligations, it is difficult to determine what is *truly* necessary and proportionate. In such circumstances, determining penalties/damages which are limited in application by territoriality may seem neither proportionate nor in compliance with norms of necessity. To avoid this dichotomy, it is crucial that criteria determining appropriate geographic scope of content restrictions be embedded in proportionality tests.

While the burden of upholding the fundamental right to expression is essentially the responsibility of the state,[46] the enforcement of content restrictions which fulfil the proportionality standard is often, and quite problematically, outsourced to intermediaries who employ techniques such as pre-publication censorship to comply with legal norms. Besides, automated tools such as upload filters are often unable to detect natural language which constitutes illegal content and can cause overly disproportionate results (Duarte & Llanso, E. 2017) that curtail fundamental rights.[47]

B. Typology of illegal content

Deciding the legality of a potentially objectionable content is a subjective exercise – content legal in one country may be illegal in another. To ensure more objectivity in dealing with

[46]See, Recommendation (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online. (2018). https://eur-lex.europa.eu/legal-content/GA/TXT/?uri=CELEX:32018H0334 (para. 36), which calls for use of automated means to expeditious remove or disable terrorist content.

[47] Research indicates that such measures often affect historically underrepresented minorities. See, Keller, d. (2018, September 4). Dolphins in the Net: Internet content filters and the Advocate General's *GlawischnigPieczek v. Facebook Ireland* opinion, *Stanford Center for Internet and Society*. https://cyberlaw.stanford.edu/files/Dolphins-in-the-Net-AG-Analysis.pdf

content restrictions, a classification can be made in the following manner:

- Content universally deemed illegal: in cases where regulatory models globally agree that content restriction is valid (e.g., child pornography),
- Content universally deemed illegal but illegality determined in national contexts: in cases where variations in the extent of illegality as well as the criteria around it differ, but there is substantive convergence on illegality itself (e.g., libel/slander),
- Content that is universally not illegal but some convergence on illegality exists: in cases where there is a lower degree of normative convergence on illegality, and
- Content that is universally not illegal and even declared legal in some regions: such as blasphemy laws.

C. Proper identification of content

Online content can be traced from varied sources including individuals (trusted notifiers or individuals voluntarily flagging third party content), private organizations (press, interest groups, civil society organizations, etc.), intermediaries (including moderators and contractors employed by such platforms), public authorities (such as judicial bodies, law enforcement agencies, or regulators) or supranational arrangements (such as the European Convention on Human Rights, multilateral treaties, etc.). As previously stated, detection of objectionable content can be a difficult exercise given the intensity of materials regularly posted online. For private sector intermediaries to act as the first line of defense, policies and mechanisms which can appropriately monitor sources and flag illegitimate content need to be set up.

Besides prevalent content flagging tools such as notice and takedown systems, keyword filters, or flagging of content deemed illegal by users, automation is another mechanism that has evolved significantly in this area. However, harmonization across intermediaries on the topic of AI moderation is nascent, particularly at the stage of identification but also in evaluating legality, determining appropriate action, and prescribing recourse.

To ensure AI identification meets the standard of normative coherence, tools used must not only analyze content substantively (by looking at keywords or images) but also according to their level of dissemination and virality.("Structuring Questions: Content moderation in relation to covid-19", 2020) Intermediaries must also be mindful of false positives and deploy human feedback to train algorithms. Currently, platforms can only permit or prohibit online content though it may be possible to have a more refined set of options which focus on appropriate geographic scope on restrictions, reducing virality of content, etc. at the intermediary level. However, it is also important to flag the importance of embedding transparency and accountability obligations in these processes, by using legal tools such as independent audits of automated detection systems, etc.

D. Due process

Finally, a framework to moderate abusive content requires the development of appropriate user notification mechanisms and due process across borders. As adjudicators of free speech, intermediaries, like courts, should be subject to rigorous rule of law standards.[48] In this respect, a high-level agreement on three principles is essential: first, reporting on content restrictions within and across stakeholders (primarily regulators and intermediaries) should be synchronized. This will encourage wider adoption and provide much-needed clarity on moderation practices and procedures. At the same time, harmonization should be accompanied by moderation transparency and information on the decision-making processes of platforms.[49] Stakeholders should also make efforts to determine consistent terminologies that can be applied irrespective of language barriers.

Second, prior to decision-making, users must be notified of the action and given an opportunity to be heard. Such procedural guarantees are fundamental to protecting citizen's rights and are in line with most national legislative frameworks (for instance, Article 14 and 21 of the Constitution of India embeds *audi alterem partem* [the right to be heard] as a core principle of natural justice). Doing so also allows the aggrieved to provide additional information to justify allegedly objectionable content, though the general principle of notification may be relaxed in certain circumstances where advance notification is impractical or not permissible, such as the threat of a terrorist attack.("User notification in online content moderation", 2020) A potential drawback in the enforcement of user notification criteria is an administrative one – large volumes of content adjudicated in a relatively short period of time may render compliance difficult. To ease this burden, intermediaries can distinguish between content that is manifestly illegal and content which requires evaluation, adopting different approaches to notification depending on the scope of illegality.[50] For instance, Germany's *NetzDG* prescribes different response times for different types of content – while manifestly unlawful content must be taken down within 24 hours, other illegal content can be taken down in a span of 7 days after receipt of complaint.[51]

Third, a platform's remediation policies, including appeal mechanisms must be clearly specified. The lack of a transparent appellate mechanism is perhaps the greatest flaw in existing content moderation practices, and requires serious consideration. While it is true that actions against platforms' decisions in national courts is possible, it is inefficient to appeal given complexities in jurisdiction issues and high transaction costs. A public-private effort to reduce economic costs to remediation and make it mainstream will prove to be an incredible effort in ensuring compliance with due process standards. At the same time, it is also important

[48] Supra n 43.

[49] While transparency reports are published by internet companies, they do not detail decision-making processes.

[50] More specifically, the timing of user notification is also crucial – for content manifestly objectionable, notification may be given simultaneously with the action, whereas for others it may precede the action as much as possible to allow some time for contextual information from users prior to making a decision.

[51] For example, German Criminal Code (Strafgesetzbuch – StGB). s. 130. https://www.gesetze-im-internet.de/englisch_stgb/englisch_stgb.html

to ensure that public-private partners are held accountable in accordance with established benchmarks for human rights and enable proper enforcement of business and human rights obligations. Efforts on this front are ongoing, including the development of e-courts in the European Union to adjudicate on the legality of content.(Ness, 2020)

8. Legal Solutions and Recommendations

Evidently, courts have had an increasing appetite for extraterritorial content restrictions in recent years. While the framework prescribed above aims to advance the goal of harmonization in adjudication practices globally, current solutions are needed to bridge the emerging divide between enforcement practices and their cross-border legality. In exploring approaches to legal solutions, overcoming existing regulatory challenges is an obvious first step. This section explores some best practices as well as technical solutions that can aid decision-making in the future.

8.1. Convergence on the terms of service of intermediaries

As noted in section five, community guidelines or terms of service framed by intermediaries to establish norms on cross-border expression in cyberspace often form the preliminary basis for content restrictions.(Lessig, 1999) Intermediaries address several aspects in their community guidelines: moderation policies, limitation of liability, and/or privacy. Given that the terms of service are contractual arrangements between the platform and the user, they directly govern the relationship between the user and the company and have a potentially higher impact on users' activity than legislation.(Bygrave, 2015)

Given their importance, specific provisions of the community guidelines which describe choice of law and/or forum can form crucial tools in addressing jurisdictional claims to content moderation. For instance, the guidelines can determine the scope of jurisdictions and establish standards against which content legality will be weighed. Contractual convergence on such provisions can significantly assist private international law regimes globally, and avoid conflict of laws situations. At the same time, self-regulation through community guidelines also assists in law enforcement and can provide much-needed clarity on treatment of objectionable content.

While it is very useful to ensure convergence on community guidelines, a crucial issue highlighted earlier in this paper needs to be revisited: how can we ensure that in formulating the terms of service, platform intermediaries will not selectively interpret human rights? What kind of benchmarking process is needed to ensure that content moderation policies are harmonized and grounded in respect for human rights?

A potential answer could lie in stakeholder-oriented model codes of conduct, such as the one formulated by the European Union on countering illegal hate speech online. (European Comissison, 2016) Though voluntary, if properly implemented, model

codes can be firmly grounded in human rights benchmarks while also providing due transparency and inclusivity for affected stakeholders. Simultaneously, we must be cautious to ensure they do not perpetuate existing issues.[52] To be clear, community guidelines may not always self-enforce (and sometimes enforce in a way which may be detrimental to the users), hence it is essential that there are proper checks at various levels. For instance, convergence on a guiding framework at the international level which focuses on the accountability of social networks in implementing community guidelines may offer additional safeguards against misuse. Another way forward is for governments to mandate disclosure policies on platforms, to ensure greater transparency on their community rules and its implementation.[53] In essence, formulating a clear and enforceable process of implementation will proscribe community guidelines from becoming a mere public relations exercise.

8.2. GeoIP filtering

One of the major obstacles in enforcing content restrictions appropriately is a determination of the physical location of internet users. Indeed, cyberspace is borderless. From a moderation perspective, this key feature creates many legal complications. For one, knowing the location of the crime may enable courts to provide targeted interventions and limit the scope of the penalty to the extent of the crime. To some extent, technologies such as geo-location can aid courts in dealing with jurisdiction questions.

GeoIP filtering allows content providers to ascertain the physical location of users by relying on GPS, WiFi information, and IP addresses. Currently, diversification of content on the internet appears to be heavily governed by geolocation technologies. (Cambridge Consultants, 2019) Indeed, several courts have emphasized the importance of geoIP filtering techniques but there remains much divergence in its adoption. For instance, while in *Yahoo! France* the court concluded that the identification of 70 percent of French residents can be estimated using IP addresses (Yahoo! Inc. v. LICRA, 2001), another Australian case completely rejected its application, holding that “no means exist to exclude transmission of material published online in any geographical area”.(Macquarie Bank Limited & Anor v Berg, 1999) Within developed jurisdictions as well, its adoption remains fragmented – United States courts have viewed access-blocking software such as geo-location as an “imperfect developing technology” that can limit the reach of content (Plixer International, Inc. v. Scrutinizer, 2018), whereas the European Union has long held the view that distribution of content online is, in principle, global. (Bolagsupplysningen OÜ and Ingrid Ilsjan v Svensk Handel, 2017)

Inevitably, the desirability of geolocation technologies reignites the debate between a

[52]For instance, the European Union Code of Conduct suffers from the similar defects as currently exist, and is therefore, limited in effect. It defines “illegal hate speech” broadly, does not provide sufficient due process guarantees or offer strong commitment to freedom of expression and has a propensity to promote ‘censorship’. For this reason, it has been strongly criticized by civil society and academics for failing to comply with international standards on the right to freedom of expression. *See,* Article 19. (2013). *Submission to the Consultations on the European Union’s justice policy.* https://www.article19.org/data/files/medialibrary/37412/A19-contribution-EU-consultations.pdf.

[53]This was highlighted recently in a report by the French government. *See,* Republique Francaise. Interim mission report. ‘*Regulation of social networks – Facebook experiment*’. Submitted to the French Secretary of State for Digital Affairs. https://bit.ly/3rblUoi

global, unrestricted internet vs. a fragmented version that results from the extensive use of such technologies.(Svantesson, 2008) However, on balance, geofiltering still remains a more realistic alternative to the global takedown orders that are currently prevalent. While its use is slowly gaining traction, technical capacity-building for law enforcement, courts, and other stakeholders, particularly in developing regions, is essential. Regulators in European Union have already taken a step in this direction by regulating geoblocking[54] and addressing its discriminatory implications in the e-commerce context. Other technical measures limiting access to content include court-ordered blocking of domain name systems, content filtering at national network levels, as well as, in extreme cases, internet and service shutdowns.

8.3. International cooperative efforts

While a no-brainer, existing tools applying multistakeholderism need to move beyond the existing realm of government-only cooperation and extend the conversation to include businesses and civil society. The lack of cross-sectoral collaboration is evident given that governments are sovereign authorities that regulate content. However, there is great merit in coordinating content moderation practices with stakeholders in order to balance consumer protection with human rights.

Some efforts on this front are worth acknowledging. In 2015, the European Union Internet Forum was set up to allow coordination of hate speech and terrorist content between governments and technology companies.("EU Internet Forum Ministerial", 2022) Bilateral engagements to handle content removal requests are also prevalent. These include the France-Germany agreement to provide content removal guidelines for intermediaries or the French-British Action Plan("French-British action plan", 2017) to regulate criminal use of the internet. Additionally, the Dynamic Coalition on Platform responsibility of the UN Internet Governance Forum has worked extensively to enhance platform responsibility and develop model contractual provisions for intermediaries.(IGF, n.d.) Other public-private initiatives such as the European Union Code of Conduct on hate speech, mentioned earlier,[55] and cooperation mechanisms between tech firms and governments to remove extremist content (Schechner, 2015) are gaining momentum as well.

It must be borne in mind that attempting to satisfy the preferences of myriad heterogeneous stakeholders can often be counterproductive, undermining decision-making processes before they are even implemented.[56] However, content moderation challenges can only be resolved if stakeholders pursue common goals by collaborating together. No nation can alone control the internet or its content. Small leaps in this direction are the beginning of the evolution of a global jurisprudence on content moderation and will require continuous efforts to iron out substantive and procedural legal inconsistencies across national legal systems.

[54] *See*, Regulation (EU) 2018/302 of the European Parliament and of the Council of 28 February 2018 on addressing unjustified geo-blocking and other forms of discrimination based on customers' nationality, place of residence or place of establishment within the internal market and amending Regulations (EC) No 2006/2004 and (EU) 2017/2394 and Directive 2009/22/EC. (2018). https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32018R0302

[55] *See,* supra 101.*P.*

[56] *See,* Christchurch Call to Eliminate Violent and Extremist Content Online (2019). https://www.christchurchcall.com/

Conclusion

This report has delved into a descriptive scoping exercise to determine approaches to content moderation in select jurisdictions. On a deeper level, it has attempted to analyze prevailing practices on both ends of the spectrum – in private and public sectors. A major thematic focus of this study is the enforcement of content moderation laws by judicial forums. Given the emerging scope of such restrictions in regions such as India and Australia, such cases have only recently come to the fore and are few in number. Nevertheless, they highlight a dangerous trend in extraterritoriality. With currently evolving principles seeking global enforcement remedies to avoid courts cementing such practices, much of this report has sought to explore the root causes of conflict and provide legal approaches to solutions.

A major import of this paper has also been to analyze the jurisdictional patchwork in applying proportionality principles to content restrictions. At the same time, it has attempted to underscore specific recommendations on new normative benchmarks and procedural mechanisms that depart from erstwhile norms of territoriality-based internet governance. By advocating for common definitions and shared typologies as well as agreed upon normative frameworks at the international level in moderating online content, this study has advanced a six-part threshold test to establish consistency in extraterritorial content restrictions practices. It has also provided a set of guiding principles to ensure effective enforcement. Findings from this study may foster a productive discussion on the issue of cross-border content moderation and inform future efforts by regulators, courts, and the private sector.

Bibliography

Journals

Brugger, W. (2002). Ban On or Protection of Hate Speech? Some Observations Based on German and American Law. *Tulane European and Civil Law Forum.* 15

Elkin-Koren, N., & Perel, M. (2020). Guarding the guardians: content moderation by online intermediaries and the rule of law. *Oxford Handbook of Online Intermediary Liability*. 18

Greenberg, M. (2003). A return to lilliput: the LICRA v. Yahoo! case and the regulation of online content in the world market. *Berkeley Tech Law Journal.* 18 Berkeley Tech. L. J. 1191 (2003). 12

Keller, D. (2019, January 29). Who do you sue? State and platform hybrid power over online speech. *Aegis Series Paper No. 1902* 6

Lessig, L. (1999). The law of the horse: What cyberlaw might teach. Harvard Law Review. 113 (506). 33

Solow-Niederman, A., Franco, F.J.C., Reventlow, N.J., & Krishnamurthy, V. (2017). *Here, there, or everywhere? Assessing the geographic scope of content takedown orders.* Working Paper— March 27, 2017 12

Svantesson, D. (2008). Geo-location technologies and other means of placing borders on the 'borderless' internet. *J. Marshall J. Computer & Info. L.* 101 (2004). 36

Svantesson, D. (2017). *Solving the internet jurisdiction puzzle*. Oxford: Oxford University Press. 10

Weitzenboeck, E. M. (2014). Hybrid net: The regulatory framework of ICANN and the DNS. *International Journal of Law and Information Technology*, 22(1), 49–73. 10

Books

Bygrave, L.A. (2015). Internet governance by contract. Oxford, United Kingdom: Oxford University Press. 33

Mill, J.S. (1963). *Collected Works of John Stuart Mill, ed. John M. Robson,* 323 vols. Toronto: University of Toronto Press, London: Routledge and Kegan Paul. 28

Ryngaert, C. (2015). *Jurisdiction in International Law.* 2nd edn. Oxford: Oxford University Press. 12

Schmitt, M. (Ed.). (2017). *Tallinn manual 2.0 on the international law applicable to cyber*

operations. Cambridge: Cambridge University Press 12

Reports

Republique Francaise. Interim mission report. '*Regulation of social networks – Facebook experiment*'. Submitted to the French Secretary of State for Digital Affairs. 35

Article 19. (2013). *Submission to the Consultations on the European Union's justice policy.* 34

De Street, A. et al. (2020). Online platforms' moderation of illegal content online: law, practices and options for reform. European Parliament. 18

French-British action plan: internet security (2017). 37

Impact of measures to address terrorism and violent extremism on civic space and the rights of civil society actors and human rights defenders - Report of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism. (2019). A/HRC/40/52. *United Nations General Assembly.* 27

Kaye, D. et al. (2019, February 1-2). Social media councils: from concept to reality. *Stanford University Global Digital Policy Incubator (GDPi).* 19

Mandates of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression; and the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism. (2019, April 4). *Office of High Commissioner for Human Rights.* 19

Ness, S. (2020, June 16). Freedom and accountability: A transatlantic framework for moderating speech online. *Annenberg Public Policy Center* 33

Pyburne, P., & Jolly, R. (2014). Australian Governments and dilemmas in filtering the Internet: juggling freedoms against potential for harm. *Parliament of Australia.* 23

Research Report by the Mandate of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, with the support of the International Justice Clinic at the University of California, Irvine School of Law. (2020, July 30). *Office of High Commissioner for Human Rights.* 18

Rosen, G. (2020, November 19). Community standards enforcement report, November 2020. *Facebook* 3

Svantesson, D. (2019). Internet and Jurisdiction Global Status Report, *Internet and Policy Jurisdiction Network.* 21

Taye, B. (2021) Shattered dreams and lost opportunities. *Access Now*. 14

Use of AI in content moderation. (2019). *Cambridge Consultants.* 35

Newspaper Articles/Blogs

Bhardwaj, D. (2021, June 7). 6k social media content takedown orders this year. *Hindustan Times* 3

Duarte, N., & Llanso, E. (2017, November 28). Mixed messages? the limits of automated social media content analysis. *Center for Democracy and Technology.* 30

Duoek, E., (2018, June 6). U.N. Special Rapporteur's Latest Report on Online Content Regulation Calls for 'Human Rights by Default'. *Lawfare.* 6

Keller, D. (2019, January 29). Who do you sue? State and platform hybrid power over online speech. *Lawfare Blog.* 22

Keller, D. (2018, September 4). Dolphins in the Net: Internet content filters and the Advocate General's GlawischnigPieczek v. Facebook Ireland opinion, *Stanford Center for Internet and Society*. 30

Koslov, V. (2015, August 10). Warner Bros. Wins Case Against Russian Internet Pirates in Landmark Ruling. 22

Namubiru, L. (2018, July 6). Uganda is making ISPs block pornography from its citizens. *Quartz Africa.* 19

Ost, F., & Kerchove, M.V. (2019). *From the pyramid to the network: for a dialectical theory of law.* Brussels: Presses de l'Université Saint-Louis 10

Paul, S. (2021, June 4). Australian media fined $840,000 for gag order breach in Pell sex assault case. *Reuters* 16

Press. (2021, April 26). India Covid: anger as twitter ordered to remove critical virus posts. *BBC.* 14

Riquelme, R., & Galindo, S.J. (2017, December 6). Richter v. Google: La Suprema Corte confirma sentencia contra Google en México. *El Economista.* 14

Russell, J. (2015, April 10). Japanese Court Orders Google To Delete Critical Reviews From Google Maps. *TechCrunch.* 22

Satariano, A. & Issac, M. (2021). *Facebook Whistle-Blower Brings Campaign to Europe After Disclosures.* New York Times. 10

Schechner, S. (2015, April 22). Tech firms, french police ally in terrorism fight. *Wall Street Journal*. 37

Schechner, S., Barrett, D. & Fowler, G. (2012). U.S. Shuts Offshore File-Share 'Locker'. *Wall Street Journal*. 24

Volokh, E. (2015, June 18). Supreme Court reaffirms broad prohibition on content-based speech restrictions. *The Washington Post.* 21

Walker, S. (2020, December 21). The fundamental basics of content moderation. *New Media Services Pty.* 8

Webpages

Alimonti, V. (2020, 7 June). New hasty attempt to tackle fake news in Brazil heavily strikes privacy and free expression. Electronic Frontier Foundation. 19

Barlow, J.P. (1996). *A declaration of the independence of cyberspace*. 12

Calster. V. G. (2017). Close, but no sigar. The CJEU on libel, internet and centre of interests in Bolagsupplysningen 16

Christchurch Call to Eliminate Violent and Extremist Content Online (2019). 37

Dynamic coalition on platform responsibility (DCPR). *Internet Governance Forum.* 37

EU Internet Forum Ministerial: towards a coordinated response to curbing terrorist and child sexual abuse content on the internet (2021, January 26). 36

GNI expresses concern about proposed 'fake news' law in Brazil. *Global Network Initiative.* 19

Keller, D. (2019). Dolphins in the Net: Internet Content Filters and the Advocate General's Glawischnig-Piesczek v. Facebook Ireland Opinion. *Stanford Centre for Internet and Society.* 16

Nojeim, G. (2014, July 30).Microsoft Ireland case: can a US warrant compel a US provider to disclose data Stored abroad?. *Centre for Democracy and Technology.* 13

Pallero, J. et al. (2019). Contributions for the democratic regulation of big platforms to ensure freedom of expression online: A Latin American perspective for content moderation processes that are compatible with international human rights standards. *OBSERVACOM*. 33

Press Release (2016). European commission and IT companies announce code of conduct on illegal online hate speech. *European Commission.* 34

Press Release. (2019). UN rights experts urge India to end communications shutdown in Kashmir. *United Nations Office of the High Commissioner for Human Rights.* 27

Schuler, M. & Znaty, B. (2020, May 21). New law to fight online hate speech to reshape notice, take down and liability rules in France. *TaylorWessing.* 20

Structuring Questions: Content moderation in relation to covid-19 (2020, April 20). *Internet and Jurisdiction Policy Network.* 31

User notification in online content moderation. (2020, November 30). *Internet and Jurisdiction Policy Network.* 32

National/International Law

Act to improve law enforcement in social networks (Network Enforcement Act - NetzDG). s. 3. 20

African Charter on Human and People's Rights. art. 9. 26

American Convention on Human Rights. art. 13. 26

Bill No. 2,630/2020 on "Freedom, responsibility and transparency on the internet (Br.). 19

Constitution of India, 1951. art. 51. 26

Convention on the Rights of Persons with Disabilities. art. 21. 26

Convention on the Rights of the Child. art. 13. 26

C*riminal Code Amendment (Sharing of Abhorrent Violent Material) Bill 2019* (Austl). 8, 19

Digital Security Act, 2018. s. 25,28-29,31 (Bangl). 19

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 ('Directive on electronic commerce'). (2000). 22

Draft Information Technology [Intermediaries Guidelines (Amendment) Rules], 2018 (Ind.). 9, 19

Electronic and Postal Communications (online content) Regulations, 2020. s. 4-5 (Tanz.). 19

European Convention on Human Rights. art. 10. 26

German Criminal Code (Strafgesetzbuch – StGB). s. 130. 32

Information Technology (Intermediaries Guidelines) Rules, 2011 (Ind.) 9

Information Technology Act, 2000. s. 69A (Ind.). 9

International Convention on the Elimination of All Forms of Racial Discrimination. art. 5. 26

International Convention on the Protection of the Rights of All Migrant Workers and Members of Their Families. art. 13. 26

International Covenant of Civil and Political Rights. art. 19(1) & 19 (2). 26

International Covenant of Civil and Political Rights. art. 19(3). 15, 27

Law no. 2020-766 of June 24, 2020 aimed at combating hateful content on the internet. art. 1 (Fr.). 19

Model Defamation Provisions (2020). *Parliamentary Counsel's Committee.* 15

Privacy Act, 1988. s. 6A. 21

Proposal for a Regulation of the European Parliament and of the Council on contestable and fair markets in the digital sector (Digital Markets Act). European Parliament. (2020). 26

Protection from Online Falsehoods and Manipulation Act, 2019. s. 7 (Sing.). 19

Recommendation (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online. (2018). 29

Regulation (EU) 2018/302 of the European Parliament and of the Council of 28 February 2018 on addressing unjustified geo-blocking and other forms of discrimination based on customers' nationality, place of residence or place of establishment within the internal market and amending Regulations (EC) No 2006/2004 and (EU) 2017/2394 and Directive 2009/22/EC. (2018). 36

Case Law

Akdeniz v. Turkey (dec.). No. 20877/10 (2014) (EU). 24

Anonymous plaintiff v. Higher Regional Court Celle (Oberlandesgericht Celle). ECLI:DE:BVerfG:2019 (2019) (EU) 24

Asociación de Gestión de Derechos Intelectuales (AGEDI) vs Neij Holdings LMT-The Pirate Bay. (2015). 24

Bolagsupplysningen OÜ and Ingrid Ilsjan v Svensk Handel AB. ECLI:EU:C:2017:766 (2017) (EU). 35

Bolagsupplysningen OÜ and Ingrid Ilsjan v Svensk Handel AB. ECLI:EU:C:2017:766 (2017). 16

Cases C-509/09 eDate Advertising GmbH and Others v X and Société MGN Limited and C-161/10 Martinez and Martinez. ECLI:EU:C:2011:685 (2011). 15

Dow Jones and Company Inc v Gutnick 210 CLR 575 (2002). 15

Dramatico Entertainment Limited v British Sky Broadcasting Limited. [2012] EWHC 268 (Ch). 24

GEMA v Deutsche Telekom. I ZR 174/14. (2014) (Ger.). 24

Glawischnig-Piesczek v. Facebook Ireland Limited, Case C-18/18 (2019). 16

Google Inc. v. Equustek Solutions Inc. 2017 SCC 34 (Can.). 22

Google Spain SL v. Agencia Española de Protección de Datos (AEPD). Case C-131/12 ECLI:EU:C:2014:317 (EU). 22

Google v. Commission nationale de l’informatique et des libertés (CNIL). Case C-507/17 (EU). (2019) 22

Haaretz.com v. Goldhar. 2018 SCC 28. 17

Handyside v. the United Kingdom. Case No. 5493/72 (1976). 26

Hassel v. Bird. 420 P.3d 776 (US) 22

In re Search Warrants Nos 16-960-M-01 and 16-1061-M to Google. 2:16-mj-01061-TJR. (2017). 9

Key Systems Case (Germany). (2014). 24

Macquarie Bank Limited & Anor v Berg. [1999] NSWSC 526 (Austl.) 35

Opinion of Advocate General Wathelet in General in Concurrence Sàrl v Samsung Electronics France SAS and Amazon Services Europe Sàrl. ECLI:EU:C:2016:843 (2016). 13

Peter Pammer v Reederei Karl Schlüter GmbH & Co. KG (C-585/08) (EU) and Hotel Alpenhof GesmbH v Oliver Heller (C-144/09). C-585/08 (2010) (EU). 36

Plixer International, Inc. v. Scrutinizer GMBH, No. 18-1195 (1st Cir. 2018) (US). 35

Ramdev vs Facebook. CS (OS) 27/ 2019 (2019). 11

Singhal v. Union of India. (2013) 12 S.C.C. 73 9

Stichting Brein v. Ziggo BV. C-61-/15 (2015) (EU). 24

Telcinco v. YouTube (2014) (Esp.) 24

Twentieth Century Fox Film Corporation And Others V Sky Uk Ltd And Others. [2015] EWHC 1082 (Ch). 24

UPC Telekabel Wien GmbH v Constantin Film Verleih GmbH and Wega Filmproduktionsgesellschaft mbH. C-314/12 (2014) (EU). 24

X v Twitter Inc [2017] NSWSC 1300 (Austl.). 23

X v Y & Z [2017] NSWSC 1214 (Austl.) 23

Yahoo!, Inc. v. La Ligue Contre Le Racisme et L'Antisemitisme, 145 F. Supp. 2d 1168,1171 (N.D. Cal. 2001) (US). 35

Yahoo!, Inc. v. La Ligue Contre Le Racisme et L'Antisemitisme, 145 F. Supp. 2d 1168,1171 (N.D. Cal. 2001). 12

YouTube, LLC v. Geeta Shroff. 2018 SCC OnLine Del 9439 (Ind.). 23

Zippo Manufacturing Co. v. Zippo Dot Com, Inc., 952 F.Supp. 1119 (W.D. Pa. 1997). 17

Annex 1

List of Expert Interview Partners

Region	Country	Expert Interviewed	Affiliation
East Asia (EA)	Japan	Yuki Mukai	Assistant Editor, Institute of Developing Economies Advanced School (IDEAS), Japan
	China and Hong Kong	Rocky Tung	Director and Head (Policy Research) Fi-nancial Services De-velopment Council,
South Asia	India	Ranjeet Rane	ReBIT | Reserve Bank Information Technol-ogy Pvt Ltd., India
		Ananth. S	Research Consultant, Reserve Bank of In-dia, India
South East Asia (SEA)	Indonesia	Ira Aprilianti	Micro Sev Financial Inclusion, Indonesia
	Philippines	Yoonee Jeong	Digital Policy, ADB, Singapore
	SEA	Cherie Teseng	COO, Secure Solu-tions Group, Singa-pore