SUMMER SCHOOL ON PLATFORM GOVERNANCE

March 20 – 26, 2023

Organised by

Centre for Communication Governance at National Law University Delhi

&

Hans-Bredow Institut, University of Hamburg

&

Institute for Technology and Society of Rio de Janeiro (ITS Rio)

at

National Law University Delhi | Sector-14 Dwarka, New Delhi

SUPPORTED BY

TABLE OF CONTENTS

INTRODUCTION

Between March 20-26, 2023 the Centre for Communication Governance at National Law University Delhi (CCG NLUD) hosted the Summer School on Platform Governance along with our partners, the Hans-Bredow Institut, University of Hamburg, the Institute for Technology and Society of Rio De Janeiro (ITS Rio), and CCG NLUD. The Summer School was supported by the Friedrich Naumann Foundation for Freedom, South Asia.

Online communication platforms have become integral to human interaction over the last two decades. With billions of users, these platforms have an enormous impact on how people communicate, consume news and information, form opinions on issues, and engage with public institutions. Governments across the world are grappling with the challenge of how to regulate platforms effectively without restricting free expression and innovation.

The goal of the week-long program was to bring together a diverse set of scholars, instructors, and students for an intensive week of lectures, discussions, and interactive classroom activities on the subject of Platform Governance.

The lecture schedule included sessions on Content Moderation, Platform Impact on Elections, The Oversight Board, Platform Transparency, The Content Safety Ecosystem, the EU's Digital Services Act, and Platforms, Competition & Consumer Protection Issues. The Summer School Program also included a keynote lecture by Shyam Divan (Senior Advocate, Supreme Court of India) and a guest lecture by Pamela San Martin, a former Electoral Councilor at the National Electoral institute in Mexico and current member of Meta's Oversight Board.

The Summer School featured twenty-eight students from Brazil, Germany, and India. As part of their involvement in the Summer School, the students were required to prepare short essays on contemporary issues within the realm of platform governance.

The essay topics cover several important themes that come within the ambit of platform governance such as, the human rights obligations of platforms, including free speech and due process, online safety, privacy and anonymity, transparency, and the role of competition law as a platform governance tool. To this end, students prepared essays on: platforms' obligations to grant their users due process when removing user content; evaluating Meta's Oversight Board as a tool for accountability; the benefits and risks of proactive monitoring technologies; the pros and cons of encryption, and an examination of the special measure taken by platforms to combat tech-facilitated gender-based violence and online extremism. Essays also covered the impact of platforms on elections, with a specific focus on false information, and the role of transparency in ensuring accountability from platforms. Given the interdisciplinary nature of the Summer School, students also prepared an essay on the use of competition law to regulate platform behaviour.

The following essays were prepared by the students of the Summer School during their time in India. We congratulate the students for their hard work in putting these essays together, and we hope our readers will find them an enjoyable and insightful read.

COMPETITION AMONGST PLATFORMS

by
Karthikeyan Murugan, Merle Heine, Vitória Oliveira

INTRODUCTION

There is little variety on the market when it comes to large online platforms, as largely a few main players share the online market among themselves. This inevitably leads to competition issues which have a negative impact on the end user experience as well as competing companies. End users often have no other alternative than using the few large online platforms in order to participate in the online world and new entrants to the market struggle to keep up with the big players. This leads to unfair competition which can be seen in the past couple of years.

To get all the competitors on a "level playing field", experts are trying to identify ways to address unfair competition and disproportionate market power. While some legislators see the necessity to regulate, other experts try to find solutions within the free market. The EU Commission, for example, believes that the concerns can only be countered with relatively strict regulation through targeted laws and corresponding supervisory authorities.

This paper discusses approaches by the EU, India and Brazil taken to address competition concerns related to large online platforms. In addition, it discusses the issue of international enforcement. Finally it will provide possible solutions to the problems of platform power through (i) regulation and (ii) the free market.

APPROACHES REGARDING COMPETITION CONCERNS

EUROPEAN UNION

While every country in the European Union has its own laws and authorities for enforcing the law, the European Commission is working on introducing the *Digital Markets Act.*[1]The *Digital Markets Act* serves to harmonize the legal framework within the EU to target large online platforms when it comes to their superior market power.

[1] REGULATION (EU) 2022/1925 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act)

The *Digital Markets Act,* which will be applicable as of the beginning of May 2023, aims to address unfair practices that result out of large online platforms such as Facebook, Google acting as "gatekeepers". A lot of new laws will be introduced in the *Digital Markets Acts* in order to support the fair market and introduce a "level playing field" without depriving these gatekeepers of the opportunities to innovate and offer new services.

The *Digital Markets Act* also introduces laws that support businesses that depend on gatekeepers (among other things). For example, these gatekeepers will have to offer interfaces in order to allow third-parties to inter-operate with the gatekeeper in specific situations.[2] Moreover, the Digital Markets Act provides business user access to data they generate when using a platform of a gatekeeper.[3]

If a gatekeeper does not comply with the laws, fines and periodic penalty payments can be introduced. The fines can be imposed in an amount of up to 10% of the company's total worldwide annual turnover, or up to 20% for companies who repeatedly violate the laws.[4] While there are some concerns regarding a negative influence on innovation, the reception of the Digital Markets Act is generally positive because of the urgent necessity to stop and prevent abuse of power due to the market dominance of gatekeepers.

INDIA

The unique characteristics of emerging digital markets have received attention in the Indian mainstream media of late due to the exponential rise in technology penetration and access to the internet in India. However, as a consequence, a fair share of the companies involved have had their brush with the Indian government[5] due to their monopolistic and restrictive trade practices they had been practising. Although there does not yet exist a dedicated specialised agency to regulate the conduct of big tech, the umbrella body in the form of Competition Commission of India has been keeping rightful vigil and has more recently seized opportunities to initiate necessary proceedings to investigate such behaviour.[6]

[2] Digital Markets Act, Recital 57.

[3] Digital Markets Act, Recital 46.

[4] Digital Markets Act, Article 30 (1) and (2).

[5] India: The Watchdog on the Trail of Big Tech. (2023), https://www.mondaq.com/india/antitrust-eu-competition-/1276498/the-watchdog-on-the-trail-of-big-tech Accessed on 24.03.2023

[6] ENS Economic Bureau, 'CCI Chairman Ashok Kumar Gupta: Digital markets lately have become 'centres for unchecked dominance'' https://indianexpress.com/article/business/cci-chairman-digital-markets-lately-have-become-centres-for-unchecked-dominance-7429170/

There has been an increasing demand to amend the Competition Act, 2002 to equip the Competition Commission of India with a dedicated branch to regulate with greater teeth the alleged inherent anti-competitive big tech market. Multiple bills have been moved to amend the Act of 2002, which was also endorsed by the Parliamentary Standing Committee on Finance (2022-2023). The Committee released its 53rd report titled 'Anti-Competitive Practices by Big Tech Companies' wherein it discussed the need to adopt an ex-ante framework to regulate monopolistic practices in the digital markets along with a recommendation to introduce an effect-based test for abuse of dominant position in the digital markets.[7]

The Committee also advised that the prospective legislation be modelled out of or be similarly placed to the Digital Markets Act of the European Union. Various government departments would be expected to work hand in hand towards realising a more holistic and comprehensive approach to this new regulatory framework.

BRAZIL

In Brazil, the competition authority ("Administrative Council for Economic Defense" or "CADE") has been operating essentially through two paths: (*i*) launching investigations of anti-competitive behaviours in digital markets; and (*ii*) publishing specialised documents on the dynamics of digital markets.

Considering (*i*), CADE has both replicated important investigations that were conducted in EU and the United States, as in *Google Shopping*, *Google Ads* and *Google Scraping*. CADE also initiated investigations on local digital players such as *iFood* - Brazil's largest platform of the food delivery market. The platform recently signed a Settlement Agreement with CADE regarding abuse of dominance due to exclusivity practices with restaurants registered on the platform, which could increase barriers to competitors in the food delivery marketplace.[8] CADE's review has been focused on unilateral conducts, which claims were consistently dismissed - which may be an indication of lack of force to battle with Big Tech.

[7] Arup Roychoudhury, 'Jayant Sinha says will table digital competition Bill on Friday' https://www.business-standard.com/article/politics/will-table-digital-competition-bill-in-parliament-on-friday-jayant-sinha-123031601277_1.html

[8]Brics Competition Centre, 'Cade And Ifood Have Signed An Agreement On Exclusive Contracts' https://bricscompetition.org/news/cade-and-ifood-have-signed-an-agreement-on-exclusive-contracts#:~:text=Brazil's%20Administrative%20Council%20for%20Economic,would%20have%20the%20same%20purpose

As for (*ii*), CADE has issued one document mapping its own decisional practice both on merger reviews and anticompetitive conducts. Additionally, CADE has commissioned a study consisting of a review of the specialised reports on digital markets around the world.[9]

Similar to India, there is pressure to update the Brazilian Competition Law (Law no. 12.529/2011), which has been in force since May 29 of 2012. One of the main arguments relies on the fact that CADE has been reviewing more than 600 transactions per year in the last couple years, but the majority of these transactions does not raise any competition concerns. However, the current criteria for transactions to trigger CADE's merger review do not necessarily encompass relevant consideration in digital markets, as they depend on the economic group's revenues of the involved parties and do not include other aspects that may be relevant in digital markets (*i.e.*, user base). In this sense, it would be a waste of public resources to review so many transactions while not capturing strategic transactions related to the digital market.

In any event, regarding legislative initiatives, two bills deserve attention: the Fake News Bill (Bill no. 2630/2020) and the Digital Platforms Bill (Bill no. 2.768/2022). The first one provides specific rules for moderation content aimed at especially large platforms, whose consumer base exceeds ten million users in Brazil, similarly to the goals of the Digital Services Act. The second one is directly influenced by the Digital Markets Act and proposes to regulate the operation of digital platforms in Brazil, assigning the National Telecommunications Agency (Anatel) as its inspection organ.

TOWARDS A REGULATED FREE MARKET

Some experts consider that the entity best placed to promote competition in markets is the government through what is referred to as market regulation. The alternative approach is through what is considered to be free market reign where only supply and demand dictate the manner in which the market functions.[10] In an ideal scenario, the free market can allow for the best interaction amongst market players and could give rise to increased efficiency and the greatest consumer utility, through 'perfect' competition.

[9]Lancieri, Filippo; Sakowski, Patricia Morita, 'Competition in Digital Markets: A Review of Expert Reports' (2020) https://www.econstor.eu/bitstream/10419/262705/1/wp303.pdf

[10]Ivestopedia, 'The Cost of Free Markets' https://www.investopedia.com/articles/economics/08/free-market-regulation.asp

It's interesting to note, however, that the interpretation of a free market appears to be changing over the years. Some scholars observe that this has resulted out of the realisation that the government is also an important player in facilitating a free market environment.[11] This occurs through regulation by the government that helps keep the market 'free' from collusion and monopoly. The case for such regulation has received greater attention in the context of digital markets.[12]

Therefore, it might best suit the need of the hour to provide regulatory direction to the market. This can further aid in facilitating the market to do what it does best and result in maximum benefit for the consumer with a healthy and competitive market environment.

The following two suggestions are offered as starting points for any new national policy towards ensuring healthy competition in digital markets.Scholars suggest that national governments put in place specialised agencies armed with the powers to investigate and proceed against suspected anti-competitive practices in digital markets before appropriate courts of law (taking example from SEC).[13] Additionally, they suggest creating a new system where data collected by the different players/companies are made available to all the players in the market.[14] This can be done by ensuring access takes place with the requisite privacy protection instruments in place.

CONCLUSION

Considering the scenario provided above, it is clear that digital platforms pose many challenges both for antitrust and overall platform regulation in a broad sense. However, regulatory responses are beginning to emerge and Europe is taking the lead on the design of ex-ante regulation. Emerging countries such as Brazil and India have been trying to learn from the European experience, as well as the American one. However, this is not a new thing - as Europe and the United States are pioneers in antitrust regulation, the Global South will be able to use these regulations as a blueprint for their own contexts.

[11] Steven K. Vogel, 'Government Regulation is the Pro-Market Solution' https://www.promarket.org/2020/10/12/government-regulation-promarket-solution/

[12] Ibid.

[13] Monti, G. (2022). Taming Digital Monopolies: A Comparative Account of the Evolution of Antitrust and Regulation in the European Union and the United States. The Antitrust Bulletin, 67(1), 40–68. https://doi.org/10.1177/0003603X211066978

[14] Ibid

After the DMA comes into force, countries from the Global South will be able to see the empirical results of ex-ante regulation. The challenge, nevertheless, is to design regulatory responses that match their own realities, considering the specifics of markets that have been recently emerging, democracies that may be weakened or institutions that may not guarantee the enforcement.

SHOULD COUNTRIES REQUIRE PLATFORMS TO REMOVE "FALSE INFORMATION" TO PROTECT THE INTEGRITY OF ELECTIONS? WHO SHOULD DECIDE ABOUT FALSE AND CORRECT INFORMATION?

by

Elder Goltzman, Pia Richter and Ananya Upadhya

INTRODUCTION

Since the coming of the post-truth world and rising internet penetration globally in general, and cases such as the Facebook-Cambridge Analytica scandal in specific, there has been a rising interest in curbing fake information online which affects electoral integrity. However, stakeholders such as governments, international organisations, social media platforms and civil society remain divided on issues such as who should act as the *"arbiter of truth"*: trained content moderators, journalists, volunteers, the user community, the judiciary, or government agencies. As the UN Special Rapporteur on free expression has noted, "false information" is vulnerable to provide authorities with excessive discretion to decide "what is truth".[1] Another issue is the extent of *intermediary liability*arising out of a failure to take down false information.

This paper analyses the problem of false information being spread on the internet in an electoral context. It deals with issues of vested interests and expediency to suggest different models on how to minimise the risks to a reliable election.

FREEDOM OF EXPRESSION AND RIGHT TO INFORMATION

Free speech is essential to keep democracy from "degenerating into tyrannies"[2]. It allows us to know our peers' ideas, confront them, and decide on important issues. We can also demand changes in the way the administration is run – or who runs it. Therefore, freedom of speech includes a collective right to information: everyone has the right to know, impart and seek information that is relevant for society as a whole, such as different political parties and their stances, especially during elections.[3]

[1] Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 2018 https://digitallibrary.un.org/record/1631686/usage?ln=en 42.

[2] Warburton, Nigel. Free Speech: A Very Short Introduction. Oxford University Press, 2009, 2.

[3]Goltzman, Elder Maia. Liberdade de Expressão e DesinformaçãoemContextoEleitorais**.** Belo Horizonte: Fórum, 2022.

In the post-truth world of politics, the rising hold of the internet as a source of information[4] means users tend to worry less about accuracy and more about sentiment. This gives room for political agents to stain their opponents' reputation or challenge the integrity of elections. Widespread false information can damage public opinion and affect voting behaviour, leading to non-representative election results.

FAKE NEWS, DISINFORMATION, MISINFORMATION AND MALINFORMATION

However, the term "fake news" lacks precision.[5] Firstly, it may include true content taken out of context. Secondly, it may not even be presented as "news", it could be content (even memes) with false information. Thirdly, the term has served a political agenda to turn citizens against the press itself. Lastly, it may also be said that news presupposes investigation and, therefore, if some content is "fake", it means it did not follow the journalistic guidelines and cannot be considered news.

Therefore, Wardle and Derakhshan's classification may be more appropriate:[6]

- Disinformation: when *false* information is knowingly shared *to cause harm* (e.g. in electoral contexts, candidates may raise doubts against the integrity of elections as an opportunity to gain support).
- Misinformation: when *false* information is shared, but *no harm is meant* (e.g. voters who truly believe a politician who says that elections are corrupt and shares it forward). This may still cause damage.
- Mal-information: when *genuine* information is shared *to cause harm*, often by "moving information designed to stay private into the public sphere" (e.g., outing of a closeted homosexual politician in order to dilute his support base).

This paper will focus on false information, i.e., misinformation and disinformation, against the integrity of elections.

[4]Puddephatt, Andrew. Freedom of Expression and the Internet**.**Cuadernos de Discusión de Comunicación e Información 6. Montevidéu: UNESCO, 2016.

[5]Tandoc, JR, Edson C,; Lim, Zheng Wei; Ling, Richard. Defining "Fake News". Digital Journalism**.** v. 6, n. 2, p. 137-153, 2017; Katsirea, Irini. "Fake news": reconsidering the value of untruthful expression in the face of regulatory uncertainty. Journal of Media Law. v. 10, n. 2, p. 159-188, 2018.

[6] Wardle, Claire; Derakhshan, Hossein. Information disorder: toward an interdisciplinary framework for research and policy making. 2017. https://rm.coe.int/information-disorder-toward-an--interdisciplinary-framework-for-researc/168076277c.

WHAT ARE OUR COUNTRIES DOING AGAINST FALSE INFORMATION IN THE CONTEXT OF AN ELECTION?

Brazil

Brazil currently has no laws dealing specifically with disinformation, but content can be restricted pursuant to a court order.[7] A bill ("PL das fake news") was not enacted since Parliament and civil society could not agree on its terms. Although there is no governmental agency specialised in disinformation, the new Lula da Silva government is organising groups to find solutions. A resolution from the Superior Electoral Court (Res. TSE nº 23.714/2022[8]) empowering itself to remove contents considered disinformation that harms integrity of national election has proved divisive. Its constitutionality was questioned in the Federal Supreme Court by the National Prosecutor's Office, but the injunction was not granted. The adjudication on merits is still pending and there is no deadline for the final judgement.

Germany

Similarly, Germany does not have a central agency for dealing with all disinformation being spread on the internet. For disseminated disinformation related to the electoral procedure in general there is a public body called the Federal Returning Officer. For identifying disinformation campaigns the liability lies with the intelligence services or other federal security agencies.[9] However, platforms are responsible for false information spreading in social media.

The Digital Services Act (DSA) clarifies the liabilities of all intermediaries who offer their services in the EU, regardless of the registered office.[10] It substantiates the notice-and-take-down procedure already regulated in two laws called 'Netzwerkdurchsetzungsgesetz' (NetzDG) and 'Telemediengesetz' (TMG). The procedure states that providers must act when they become aware of an infringement of rights. In a specified period, the platform has

[7] Marco Civil Law of the Internet, Arts 18-19.

[8] Full text in Portuguese at https://www.tse.jus.br/legislacao/compilada/res/2022/resolucao-no-23-714-de-20-de-outubro-de-2022.

[9] The Federal Returning Officer. "Identifying and combating disinformation". https://www.bundeswahlleiter.de/en/bundestagswahlen/2021/fakten-fakenews.html.

[10] Nomos. "New obligations for digital services". https://www.nomos.de/digitalrecht-hofmann-raue/.

toreview whether an infringement has occurred in a particular post and to take it down if so. The DSA also provides users with enforceable rights against platforms.[11]

India

While the Information Technology Act follows the "safe harbour" approach, the Information Technology Rules, 2021 oblige intermediaries to undertake reasonable efforts to prevent their users from uploading content which is defamatory, libellous, illegal, impersonatory, deceptive or misleading or even "patently false or misleading in nature but may reasonably be perceived as a fact" (Rule 3(1)(b)).[12] This may lead to overbroad regulation on satire. In some of these cases, the takedown requests must be processed within 24 hours (Rule 3(2)(b)). These have been criticised as potentially unconstitutional.[13]

For political ads, India has an elaborate mechanism: Each ad must obtain approval from an "Electronic Media Monitoring Committee"[14]. This is similar to South Africa's model where all official ads must be uploaded to a repository to help distinguish them from fake ads[15].

CAN PLATFORMS MODERATE? SHOULD PLATFORMS MODERATE?

It is known that platforms, being business corporations, have a vested interest in profitability and, therefore, in the "clickworthiness" of posts, especially with headlines that are lurid or shocking (and often distortions of the truth). Can they be trusted to self-regulate or even hire content moderators (third-party fact checkers)[16] to take down false information or should there be independent regulators?

[11] Nomos. "New obligations for digital services". https://www.nomos.de/digitalrecht-hofmann-raue/.

[12] Rule 3(1)(b). https://mib.gov.in/sites/default/files/IT%28Intermediary%20Guidelines%20and%20Digital%20Media%20Ethics%20Code%29%20Rules%2C%202021%20English.pdf. .

[13] Internet Freedom Foundation. "Deep dive : How the intermediaries rules are anti-democratic and unconstitutional". (Feb. 27, 2021) https://internetfreedom.in/intermediaries-rules-2021/.

[14] Marsden, Chris; Brown, Ian; Veale, Michael. "Responding to Disinformation, Ten Recommendations for Regulatory Action and Forbearance" in "Regulating Big Tech: Policy Responses to Digital Dominance" Oxford Academic.

[15] Marsden, Brown and Veale. "Responding to Disinformation".

[16] Greenberg, Andy. 'Watch Workers Learn How to Filter Obscene and Violent Photos From Dating Sites' Wired: https://www.wired.com/2017/04/watch-people-learn-filter-awfulness-dating-sites/.

The UN Special Rapporteur has noted that restriction on free speech must involve the oversight of independent judicial authorities.[17] Brazil's intermediary liability regime requires a court order to restrict particular content,[18] while India has a "notice and takedown" process that requires the order of a court or appropriate government agency.

The Special Rapporteur criticises the imposition of obligations on platforms to take down content without judicial orders, that too with heavy fines (such as Germany's NetzDG regime which requires platforms to take down "clearly illegal" content within 24 hours or face substantial penalties).[19] The pressure put on platforms, whose key interests are economic (protecting themselves from fines or blockages) may result in overbroad regulation, that is, the removal of lawful content.

BALANCING EXPEDIENCY AND DUE PROCESS

The above points may make it seem like the requirement of a prior judicial order (in each jurisdiction) may be most suited for targeted takedown of content. However, the sheer numbers of social media posts during elections – not in the least perpetrated by "troll factories" – has led to governments adopting intermediary liability rules rather than pursue the anonymous users themselves since they lack technical capacity.[20]

Another difficulty is how imminent the risks from disinformation are in the electoral context. Therefore, a "notice and takedown" model (though not supported by the Manila Principles even if the uploader has the right to appeal since the incentive structure is still weighted towards takedowns) may be a fair balance. The two models supported by the Principles, i.e. "notice and notice" may prove completely ineffectual in disinformation since the uploader intends to mislead voters anyway. "Notice and judicial takedown" might be too slow for electoral disinformation.

Meta explicitly states that it "cannot become the arbiter of truth", instead relying on its users to report disinformation and working with "independent" third-party fact-checkers to flag

[17]Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 2018 https://digitallibrary.un.org/record/1631686/usage?ln=en

[18] Marco Civil Law of the Internet, Arts 18-19.

[19]Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 2018 https://digitallibrary.un.org/record/1631686/usage?ln=en

[20]Marsden, Brown and Veale. "Responding to Disinformation".

reported stories as disputed, placing reported and disputed stories lower on the News Feed.[21] This model could be more widely adopted by other social media platforms, with moderators discovering false content to be taken down. Here, it would be imperative that since false content is not as easily identifiable or objective as pornographic or violent content (and requires cross-checking), those short deadlines cannot be adopted here.[22] For example, the Australian Communications and Media Authority specifies certain risks from disinformation which are imminent (including threats to electoral integrity),[23] which can have specialised and targeted moderation.

AUTOMATED CONTENT MODERATION?

At the same time, large social media platforms do not rely solely on human intervention to combat disinformation. Meta and Twitter have deployed AI tools at large scale claiming it to be the only cost-effective response. The use of AI works especially well in certain areas like child pornography, terrorist videos and intellectual property (where technology such as comparing hash values can allow for unlawful content to be immediately matched and spotted), but may prove ineffective in complex areas such as disinformation and result in overbroad censorship.[24] Language learning models, if unable to find the "truth" against which to judge content, will be helpless in determining what is false. Human review, therefore, becomes crucial.

SOLUTIONS

Following from the above discussion, a suitable approach may be to require platforms to hire content moderators (or work with fact-checking organisations) to decide cases by a specified process (cross-checking with a specified list of trusted news sources). Content reported by users as "false information" should be routed to these moderators and taken down if found to be false information.

---

[21] Meta for Media, "Working to Stop Misinformation and False News". https://www.facebook.com/formedia/blog/working-to-stop-misinformation-and-false-news.

[22] Greenberg, Andy. 'Watch Workers Learn How to Filter Obscene and Violent Photos From Dating Sites' Wired: https://www.wired.com/2017/04/watch-people-learn-filter-awfulness-dating-sites/.

[23] Australian Media and Communications Authority. Misinformation and news quality on digital platforms. https://www.acma.gov.au/sites/default/files/2020-06/Misinformation%20and%20news%20quality%20position%20paper.pdf.

[24] Marsden, Chris; Meyer, Trisha. 'How Can the Law Regulate Removal of. Fake News?' Society for Computers and Law. 2019.

Here, it becomes imperative to mention three points: Firstly, there must be due process in takedowns (due diligence, transparency in logging, and proportionality in the punitive measures). However, transparency measures such as revealing the identity of the complainant to the original poster might be avoided since it would create a chilling effect against those wishing to report content by those who are connected to state power. Secondly, it is necessary to provide training and support to content moderators. The case of *Selena Scola v Facebook* could be instructive in this regard.[25] Thirdly, international human rights frameworks should be used as a foundation for social media platforms owing to higher objectivity.[26]

A downside may be that satirical content might be struck unfairly due to an inability on content moderators to adjudge context. One counter could be for platforms to create a common list of producers who routinely create satirical content in good faith (for instance, *The Onion*). It must also be noted that in sensitive times such as during an election, striking down satirical content may still be a reasonable balance since it could affect voting behaviour irreversibly, even if unintentionally. For cases falling outside of this, and others where "what is truth" slowly unfolds over time, remedial measures to reinstate content are required, such as with judicial intervention (keeping in line with the Manila Principles).[27] Online appeal procedures could be processed by a special court or fast-track tribunal set up under the election management body since they are impartial and better streamlined. These tribunals could consist of election officials (who are expected to be impartial and independent), judicial officers (serving or retired), human rights experts and fact-checking entities, and not members of the political executive. As the UN Special Rapporteur states, the government should not become the arbiter of truth.[28]

Apart from citizens, elections pose the mischief of false information spread by political parties. The Indian/South African model (detailed above) can be adopted to counter these.

Additionally, it is important to not only look towards punitive measures and takedowns/ deplatforming (which may harm freedom of speech, especially owing to automation). Instead,

[25] *Selena Scola, et al. v. Facebook, Inc*. Superior Court of the State of California, County of San Mateo. https://contentmoderatorsettlement.com/.

[26] Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 2018 https://digitallibrary.un.org/record/1631686/usage?ln=en 46.

[27]Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 2018 https://digitallibrary.un.org/record/1631686/usage?ln=en 59.

[28] Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 2018 https://digitallibrary.un.org/record/1631686/usage?ln=en 68.

measures boosting transparency, like allowing users greater control over their feeds, displaying all sources of advertising (as suggested by the new EU Code[29]) and deprioritising reported content (or displaying a warning) may prove to strike the right balance.

CONCLUSION

There is agreement on the dangers to democracy caused by false information as well as on the fact that the government, composed of elected officials, cannot be the "arbiter of truth". Instead, the process of identifying and combating false information could involve content moderators and fact-checkers who are either external to the platform or hired by it, but trained in either case. In an electoral environment, when this proves ineffective, neutral special courts or fast-track tribunals could be set up under electoral and judicial officials and other experts.

[29] European Commission. The 2022 Code of Practice on Disinformation. https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation.

ENCRYPTION RIGHTS AND ITS POLICY CONCERNS

By

Aditya, Diego André Cerqueira, Eva-Maria Laas, and Heloísa Helena Silva

I. INTRODUCTION

Data encryption[1]is a hotly debated issue in much of the world today. Many big tech companies have been adopting encryption tools for their operations. This has really democratized the digital media setup as it provides the right to express freely without any elaborate costs attached to it, which was not possible earlier. Encryption also provides security and anonymity to express views freely without fear of being identified. Such protection can be especially important for vulnerable groups and minorities, ensuring their right to communicate without fear of political persecution. Furthermore, it is also relevant for investigative journalists who need to rely on anonymous sources to expose the wrongdoing of powerful state and private actors.

At the same time, governments are criticizing these encryption-based communication models for perpetuating crime and undermining national security by providing a safe passage to criminal activities. Investigating agencies claim that encryption makes it impossible for them to intercept any illegal activity. They therefore desire some check over these encryption models, such as some form of backdoor entry.

This essay will discuss the encryption related policies of three countries - India, Brazil and Germany. It will analyse the positive impact of encryption models over digital communications and also the concerns arising out of the widespread use of encryption. In the process of doing so, it will formulate solutions that preserve encryption while tackling the menace of cybercrime.

[1] Encryption is defined as the Cryptographic transformation of data (called "plaintext") into a form (called "ciphertext") that conceals the data's original meaning to prevent it from being known or used. Computer Security Resource Center, NSIT.

https://csrc.nist.gov/glossary/term/encryption#:~:text=Cryptographic%20transformation%20of%20data%20(called,from%20being%20known%20or%20used. Last accessed 6 March, 2023.

II. REGULATION PERSPECTIVES: INDIA, BRAZIL, AND GERMANY

II.1. India

In 2015, an expert committee was constituted by the union government. It recommended a National Encryption Policy,[2]to regulate the domestic use of encryption technologies, to be enacted under the Information Technology Act, of 2000.[3]However, this policy was met with severe criticism, due to its impractical stipulations such as key-size limits, and plain text retention requirements for both users and businesses.[4]The draft policy was soon withdrawn as a result and to date, no subsequent draft has been released.[5]

The Indian Supreme Court in its *K.S Puttaswamy* judgment[6]has upheld the right to privacy as a fundamental right to life under article 21 of the Indian Constitution. The court has laid down four principles under which such rights can be taken away.[7]A Committee of Experts led by Justice BN Srikrishna was established by the government in the wake of the *Puttaswamy* ruling, and it formulated its report and a preliminary version of the Personal Data Protection ("PDP") Bill in 2018.

The committee noted that the security of people's personal data is threatened by the existing low encryption standards mandated by law.[8]The committee report also referred to the safeguards present in some other jurisdictions like public transparency and legislative and judicial oversight.[9]Instead of taking a shortcut that is at odds with the principles of civil

[2] Draft Encryption Policy (2020). https://netzpolitik.org/wp-upload/draft-Encryption-Policyv1.pdf.

[3] Section 84A. IT Act, 2000.

[4] Pratik Prakash Dixit, Conceptualising Interaction between Cryptography and Law 11 NUJS L Rev 327 (2018).

[5] Sheela Bhat, Draft National Encryption Policy Withdrawn: Narendra Modi Government's Flip Flop Style, The Indian Express (2016). https://indianexpress.com/article/explained/encryption-draft-withdrawn-modi-governments-flip-flop-style/.

[6] (2017) 10 SCC 1.

[7] I. Express Legal Provision, II. Legitimate Aim of such encroachment, III. Test of Proportionality and IV. Following Due Procedure.

[8] Committee of Experts under the Chairmanship of Justice B.N. Srikrishna, A Freeand Fair Digital Economy Protecting Privacy, Empowering Indians 125 (2018).

[9] For instance, In Germany, there is legislative oversight in form of the G-10 Commission, consisting of four members appointed by the German Federal Parliament, is responsible for approving surveillance measures by intelligence agencies. Australia, US and Canada subscribe to the judicial safeguard model, in which warrants, subpoenas and other court orders are required to access content of messages in transit and in storage. See, National Institute of Public Finance and Policy, Use of personal data by intelligence and law enforcement agencies. (2018). [10]Brazilian Federal Constitution - 1998. Available at

<https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm>

liberties and national security, the state should look for viable alternatives and better implementation of the standard procedure of investigation.

II.2. Brazil

Despite the country having made significant advances in the legal framework of the Internet, through laws like Internet bill of rights (Marco Civil da Internet - 2014) and Brazilian Data Protection Law - (13.708/2018) or LGPD, at the moment, there are no specifical controls to the use of encryption in Brazil.

Nonetheless, the Brazilian constitution (1988)[10] at the same time protects freedom of expression and privacy and also forbids anonymity (Art. 5º, IV).[10]A hypothetical regulation about encryption in the country should balance those aspects, protecting free speech but ensuring a way to identify the potential authors of harmful uses of that freedom.

Since anonymity is forbidden, there are even episodes of blocking WhatsApp in the country by judicial decisions against encryption in private messaging services, as a way to force the company to comply with demands for data.[11]In those episodes and during elections, alternative messaging applications such as Telegram have been increasing between Brazilians users.[12]

II.3. Germany

In Germany, the content of a communication is protected via the constitution by several fundamental rights (Art. 10 (1); Art. 2 (1), 1 (1) GG). These rights are not only of defensive nature; they legally bind the state to protect its citizens against safety risks regarding their communication.[14] Even though the current government has stated that one of its goals is to

[10] Art. 5 Everyone is equal before the law, without distinction of any kind, guaranteeing Brazilians and foreigners residing in the country the inviolability of the right to life, liberty, equality, security and property, in the following terms: (...) IV - the expression of thought is free, anonymity being prohibited; (...) (free translation)

[11] Available at <https://www.theguardian.com/world/2016/jul/19/whatsapp-ban-brazil-facebook>.

[12] Available at <https://www.france24.com/en/live-news/20220223-brazil-s-bolsonaro-turns-to-telegram-as-vote-nears>.

[14] RalfPoscher/Katrin Kappler, Staatstrojaner für Nachrichtendienst – ZurEinführung der Quellen-Telekommunikationsüberwachungim Artikel 10-Gesetz, VerfBlog, 2021/7/06. https://verfassungsblog.de/staatstrojaner-nachrichtendienste/.

implement a "right to encryption", the past government has established some new laws to enable access to encrypted digital communication under certain circumstances.[13]

The German criminal procedure law enables the police to observe communication before or after encryption or to even decrypt encrypted content (cf. § 100a (1) StPO; § 51 (2) BKAG). The execution of these actions is only legitimate if the authority suspects a serious offense, like terrorism or organized crime. Furthermore, the police need a judicial order (cf. § 100e (1) StPO). Despite these checks and balances, fundamental rights are still negatively affected.[16]

In May 2021 the government made amendments to the existing acts regarding the protection of the constitution which enables the surveillance of encrypted communication by intelligence offices (cf. §§ 2 (1a), 11 (1a) G 10). These competencies are only applicable in cases of serious crimes as well (cf. § 3 (1) G 10). On top of that, there is legislative oversight in the form of the G-10 Commission, consisting of four members appointed by the German Federal Parliament, who need to approve these surveillance measures (cf. §§ 14 ff. G 10).[14]These amendments were heavily criticized by Data Protection organizations and legal experts; it was argued that the state was actively creating security gaps and therefore infringing its constitutional obligation to protect data.[18]

Even more critical are the current plans of the European Union to legally bypass data encryption. In May 2022, the European Commission proposed a regulation for laying down rules to prevent and combat child sexual abuse.[15]Critics fear that this regulation might put an end to end to end-encryption because it obligates online services to detect, report, remove, and block known and new child sexual abuse material, as well as solicitation of children,

[13] Available at

<https://www.spd.de/fileadmin/Dokumente/Koalitionsvertrag/Koalitionsvertrag_2021-2025.pdf>.

[16] Ralf Poscher/Katrin Kappler, Staatstrojaner für Nachrichtendienst - ZurEinführung der Quellen-Telekommunikationsüberwachungim Artikel 10-Gesetz, VerfBlog, 2021/7/06. https://verfassungsblog.de/staatstrojaner-nachrichtendienste/.

[14] Federal Commissioner for Data Protection and Freedom of Information, Telecommunications surveillance in Germany.

https://www.bfdi.bund.de/DE/Buerger/Inhalte/Nachrichtendienste/Telekommunikationsueberwachung.html [18] Ralf Poscher/Katrin Kappler, Staatstrojaner für Nachrichtendienst - ZurEinführung der Quellen-Telekommunikationsüberwachungim Artikel 10-Gesetz, VerfBlog, 2021/7/06. https://verfassungsblog.de/staatstrojaner-nachrichtendienste/.

[15]https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52022PC0209&from=EN.

"regardless of the technology used in the online exchanges" (cf. recital 26 of COM(2022) 209 final).[16]

III. OUR SOLUTION - HOW TO BALANCE COMPETING INTERESTS

Security has been a long-term concern over digital media platforms whether we take up the case of highly sensitive issues such as F-35 documents leak[17]or day-to-day privacy concerns like interception of Zoom calls.[18]

Law enforcement organizations have claimed that encryption poses a serious hazard that makes it easier for adversaries to carry out their crimes. These organizations thus say that they are unable to gather information to stop or punish offenders. There is no doubt about the usefulness of encryption technology - where we get double protection of public and private key-based systems which makes any interception or hampering with data virtually impossible. But as we know, every coin has two sides. The other side here, articulated by investigating and prosecuting authorities, is the ill-use of this technology by anti-social elements. Consequently, there is a growing consensus among national governments to make such communication platforms liable to share such information as and when required by creating some back-door entry.

But there are several major issues with such extraordinary interception. One, such a backdoor entry model will create vulnerabilities in the encryption model. It is only a matter of time before such vulnerabilities are discovered and exploited by hostile third parties. This will only hurt law abiding citizens who are forced to use less secure unencrypted/backdoored platforms for no fault of their own. It will break users' trust in platforms due to heightened privacy violation concerns.

Secondly, once backdoors are instituted, criminals currently relying on mainstream encrypted platforms will simply switch to illegitimate encrypted platforms that ignore the legal mandate

[16] Bits of Freedom, The European Commission might put a stop to end-to-end encryption, 2022/03/23. https://edri.org/our-work/the-european-commission-might-put-a-stop-to-end-to-end-encryption/.

[17] Mike O'Brien, "Pentagon Admits F-35 Data Theft is a Major Problem," Institute for Defense and Government Advancement. https://www.idga.org/archived-content/news/pentagon-admits-f-35-data-theft-is-amajor-problem. Last accessed 6 March, 2023.

[18] Kari Paul, Zoom Will Provide End-To-End Encryption To All Users After Privacy Backlash, The Guardian, (2020). https://www.theguardian.com/technology/2020/jun/17/zoom-encryption-free-calls. Last accessed 6 March, 2023.

to institute backdoors. Thus, we cannot use encryption backdoors as a quick cut to combat internet-enabled crime. The fresh dangers to security of communication platforms caused by the use of backdoors will be too significant while the benefits will be doubtful and dubious.

The third problem will be the infringement of the right to privacy by the state itself. The state is responsible for protecting the civil and political rights of individuals but at times it becomes the principal encroacher of our civil liberties because certain states go after its critics/political dissidents. Thus, encryption can be seen as a way to protect human rights in authoritarian contexts. Thus, compromising encryption through backdoors will compromise the human rights of dissidents and critics.

Nonetheless, there is a way out. To be specific there are a number of ways out.

Law enforcement agencies have been able to successfully tackle problems like terrorist content without compromising end to end encryption. They have done so by relying on investigative tools and techniques that involve cross-examining data, crossing information with internet service providers, and with the collection and harvesting of metadata to support their work. Other measures may involve relying on techniques such as user reporting. Here, users of encrypted platforms flag abusive content, which is forwarded to the platform for further investigation.

IV. CONCLUSION

Technology cannot be blamed for the ills that plague us. Take electricity for example. Today it is part of our daily lives. Indeed, this technology is essential for supporting and enabling human beings in multiple ways. But electricity, like cryptography, can be used as a weapon to intentionally cause damage. And in both cases, the technology is not the one to blame, but its use for harm and how it can be weaponized.

At first sight, creating backdoors may sound like a solution to the problem of cybercrime. But the domino effect of such a measure will be significant. Instituting backdoors could affect not only the right to privacy but democracy itself. In order to secure a way into encrypted systems we would make the whole chain of information vulnerable, possibly creating unprecedented changes in how users/citizens experience the internet in the 21st century.

Instead relying on metadata and measures such as user reporting of abusive content will enable governments to clamp down on wrongdoing while preserving privacy and civil liberties.

ONLINE CONTENT RESTRICTION AND GRIEVANCE REDRESSAL

by

Harshwardhan Pushkin Sharma, Isabela Maiolino and Marlene Spaude

1. INTRODUCTION

Social life and discussion are increasingly taking place on online platforms. Some platform interactions and information sharing result in disagreements. Platforms consequently remove content. Aggrieved users are then unsure about how to contest such removals.

This paper will consider the appropriate grievance redressal process that should be instituted to empower and enable users to challenge content takedowns by platforms. We will also look at what laws can establish standards for a user's right to a hearing before the platform or an independent body when their content is restricted.

2. PLATFORMS AS ADJUDICATORS

Platforms have come a long way from their humble beginnings—the small startups of the past are corporate behemoths of the present. Presently, they host millions of people worldwide and have multiple revenue streams. A large population with diverse opinions and values leads to numerous disagreements. These disputes necessitate dispute resolution mechanisms, most of which are formed by the platforms themselves.

In this essay we contend that the dispute resolution process developed by the platforms should be subject to certain guardrails. Specifically, we recommend the incorporation of international law standards into the terms of service of platforms. These standards can be rooted in the International Covenant on Civil and Political Rights (ICCPR) and the Universal Declaration on the Independence of Justice. Moreover we recommend the creation of independent dispute resolution bodies that are free of the undue influence of platforms. Finally we recommend the incorporation of procedural safeguards into the hearing process so that users' due process rights are respected.

3. MINIMUM STANDARDS FOR DISPUTE RESOLUTION: PRINCIPLES OF INTERNATIONAL LAW

Content moderation by digital platforms raises several concerns. As a profit-driven enterprise, it is important for platforms to retain their autonomy. At the same time, the wide-

ranging impact of digital platforms on the everyday lives of people around the world calls for more checks and balances on their powers.

Therefore, the question arises as to which standards should be applicable to decide whether content should be blocked or not. After all, on the large social platforms, such as Facebook or YouTube, almost every nation can participate in public discussions. However, different countries have different perspectives on freedom of expression.[1] Moreover, the meanings of memes and hashtags, for example, sometimes change overnight. However, the guidelines for platforms should have a certain stability and be universal.

We contend that international law standards should be applied. Inclusion of international law norms by internet platforms in their terms of service will provide greater clarity over which laws are applicable and reduce disputes. This may lead to the creation of broader, more universal standards by which states and platforms could be bound when considering human rights.

Furthermore, this will lead potentially to more transparency and increase awareness among platform users. Transparency about company policies and infotainment advertising of rules will ensure that platform users comply with the laws. Some of the norms that should be applicable include the International Covenant on Civil and Political Rights (ICCPR) and the Universal Declaration on the Independence of Justice adopted in Montreal in 1983.

The ICCPR has been signed by several states and is almost universally applicable. The right to freedom of expression is broad and includes the freedom "to seek, receive, and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, through the arts, or by any other means of his choice," according to Article 19(2) of the International Covenant on Civil and Political Rights.

Article 19(3) of the ICCPR also specifies that any restriction on a person's right to freedom of expression must meet the threefold test of lawfulness, legitimacy, and necessity. International law interpretations may borrow from the interpretation of these terms by the ECHR. The ECHR interprets it as a "living instrument." As a result, the interpretation is always adapting to new social or technological changes. It may be that a case is judged differently today than it was 10 years ago, even though the same law is used. Such an interpretation internationally

[1] See Evelyn Douek, 'Content Moderation as Systems Thinking' 136 Harvard Law Review p. 554 f.

can aid dispute resolution in cyberspace and allow space for laws to keep up with changing dimensions of the internet.

4. INDEPENDENCE OF THE DISPUTE RESOLUTION BODIES

An important concern when it comes to setting up dispute resolution mechanisms online is whether the dispute resolution bodies should be under the control of the platform itself or whether they should be independent of the platform.

We are in favour of independent dispute resolution bodies. This is because the dispute resolution body's financial reliance on the platform could lead to the oversight body being influenced in its decisions. In addition, the independent body should also be effective enough to really influence and question the platform's policies.[2] It should therefore be as autonomous as possible from the platform.

Some have suggested government oversight as an alternative to platform controlled dispute resolution bodies. But government oversight could prove difficult. As large platforms connect users globally, a multitude of states become connected. These states have different laws, which may also be influenced by the culture or religion of the country in question. If a dispute now arises between a user and the platform, this can lead to jurisdictional disputes and litigation in multiple forums.

Additionally, there should be agreement on which cases will be heard by the dispute resolution bodies. Patently illegal content, such as child abuse, pornography, or violent acts must be removed promptly - so they should not be subject to hearings. Other categories of content can be heard by the dispute resolution bodies. In this vein, it seems reasonable to prioritise content that reaches a large number of people. Particularly topical issues should also be reviewed as a matter of priority, since this information is of particular concern to users at the time.

5. PROCEDURAL SAFEGUARDSAND DUE PROCESS RIGHTS

In many cases, it is hard to decide whether content should stay online or not. For example, misinformation on platforms tends to rise around the time that elections take place.

[2] See Flynn Coleman, Brandie Nonnecke and Elizabeth M Renieris, 'The Promise and Pitfalls of the Facebook Oversight Board', 2021, CarrCenter Discussion Paper Series, p.1.

Moreover, content on the internet can have varied interpretations. Content that is posted on the internet is not always construed as the originator of content intended to be since they can be devoid of tones, gestures, and facial expressions. A good way to go is to have a responsible mechanism for both the content creators and the platforms themselves.

Platforms should be transparent about their content moderation policies and provide users with clear guidelines on what is and is not allowed on their platforms. This can help prevent confusion and ensure that users are not caught off guard when their content is removed. A good way to go about deciphering the exact meanings of uploaded content is to request users to tag or provide an explanatory note for their content prior to posting. If a software detects any content to be violative of the platform's rules, a better approach may be to not remove the content altogether but put it under consideration for further review and give the user a chance to explain the content.

So, where content can be interpreted in multiple ways and does not explicitly display nudity or violence, such content can be forwarded to the independent body. If it falls under interpretive content and doesn't outwardly show nudity or violence but a platform restricts such content, it can be forwarded to the independent body.

Where the independent body decides to hold a hearing, certain criteria should be followed. Platform users should be informed prior to the hearing as to why the content in question was restricted. Importantly, the notification sent to the user shall be in a language the user understands. An average user should be able to understand the reasons why the content was removed. In addition, the user should be given adequate time and information to examine the allegations. After all, the user should also have the time and opportunity to seek legal advice. The independent body should explain in writing why the right of expression is being restricted.

6. CONCLUSION

Large online platforms are a staple feature of our lives today. When these platforms remove content, users often have no recourse. A comprehensive dispute resolution process is therefore essential to protect the right and ability of users to express themselves freely online.

To this end we recommend several measures. International laws have been examined in order to determine possible minimum standards for dispute resolution. They should be incorporated into the terms of service of platforms. Concerns about the independence of such dispute resolution mechanisms have also been addressed. We have also discussed the importance of incorporating due process protections into the hearing process in the form of certain procedural safeguards.

In conclusion, every internet user should have the right to a hearing before the platform or an independent body when their content is restricted. The minimum requirements for this right are not too difficult to incorporate. It is hoped that such minimum standards are drafted and enforced by all stakeholders soon.

OVERSIGHT BOARD: A REAL SOLUTION OR AN OFF-STAGE PROBLEM TO PLATFORM ACCOUNTABILITY?

by

Juliana Fonteles da Silveira, Anamta Khan, and Ella Abry

1. GENERAL CONSIDERATIONS

The concentration of a few private online platforms in the intermediation of public debate and their role in controlling speech[1] challenge the optimistic aspirations of promoting a democratic culture on the internet.[2] Through the microtargeting of political ads, the recommendation of content, and a system of internal and public rules that inform content ranking and demotion, Meta governs to a great extent the flow of information online. The impacts of such practices on human rights and the democratic debate have caused strong reactions in civil society, academia, and governments to address these issues through the regulation of platforms.[3]

In this scenario, the Oversight Board represents a self-regulation movement of Meta to respond to the demands for public scrutiny, transparency and accountability of its platforms. The Board was invented by Meta CEO Mark Zuckerberg and has been operating since October 22nd 2020.

The idea was to create an independent organization, which would confirm or reverse problematic content moderation decisions on account suspension and content removal, and publish its reasoning. That is the reason why Meta's CEO called the body the company's Supreme Court.

Its task is to focus on Meta's content moderation system through case decisions and the delivery of Policy Advisory Opinions (PAOs). Despite being structured to establish a separation of powers and an oversight function in content moderation decision-making, the Oversight Board has not fully achieved its intended goals. Nor has it realized the broader

[1] Kate Klonick, 'The New Governors: The People, Rules, and Processes Governing Online Speech' (2018) 131 Harvard Law Review 1598; Jack M Balkin, 'Free Speech Is a Triangle' (2018) 118 Columbia Law Review 2011.

[2] Jack M Balkin, 'Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society' (2004) 79 Ney York University Law Review 1.

[3] Mårten Schultz. Six Problems with Facebook's Oversight Board. Not enough contract law, too much human rights

public demands for platform accountability. In this sense, this paper is dedicated to discussing the following challenges for platform accountability and Meta: i) independence, ii) scope, iii) effectiveness and iv) purpose. Finally this paper will provide some suggestions about alternatives to the existing Oversight Board.

2. PURPOSE

As these issues on platform accountability seem to remain unsolved with the implementation of the Oversight Board, an obvious question arises: Why does it exist at all? Promoting mechanisms of apparently external review of the company's decision builds trust among users and legitimates Meta's practices in addressing harms generated in the enforcement of its content policies, as well as in tackling speech that has been understood as harmful or undesirable. As users see that the company is reviewing its decisions where it has been criticized on human rights and democracy grounds, the public message that prevails is that these issues and users' safety are a priority for Meta.

While the Oversight Board has not been able to fix the origins of the problem of harmful communications on its platforms, it does perform a powerful symbolic role in validating Meta's products and services and its value of corporate responsibility. This legitimacy and trust also has the effect of maintaining users in the platform and bringing new ones, which increases platforms revenue. Therefore, the purpose of the body does not seem to be promoting meaningful platform accountability, but rather elevating profit.

3. CHALLENGES AND RISKS OF OVERSIGHT BOARD TO PLATFORM ACCOUNTABILITY

a) Independence of the body

Having been constituted by the company subject to supervision and financed by its resources – with an initial contribution of $130 million grant,[4] - the operation of the Oversight Board can indicate a lack of independence. It is worth noting that even though its advisors and work teams have a professional background in human rights, constitutional law and freedom of expression, their decisions and opinions, even those that point to flaws in the platform, rarely critically question the platform business model and its recommendation algorithms based on the exploitation of personal data and increased engagement. Decisions are restricted to discussing content and enforcement policies in individual cases that do not compromise the

[4] See Inside Meta's Oversight Board: 2 Years of Pushing Limits. Wired. 08.11. 2022. Steven Levy. Available in: https://www.wired.com/story/inside-metas-oversight-board-two-years-of-pushing-limits/

moderation system as a whole or the platform system, in a way that captures public attention for individual content moderation problems, building the concept that if they are corrected the system will work perfectly.[5]

In this order of ideas, it can be observed that the Board assumes a role of supporting and legitimizing the functioning of the platform, attributing more credibility to it,[6] insofar as: (i) carries out the task of addressing individual errors or risks to fundamental rights - arising from the platform design itself and its products and services -, (ii) has hired specialists from various sectors who are especially concerned and dedicated to the protection of human rights and are conveniently well connected with those actors who scrutinize and demand solutions from the platforms, (iii) offer some degree of transparency, considering that it publicly discloses its decisions. This configuration contributes to shape an image of impartiality, independence, and accountability of decisions about the platform's content moderation practices.[7]

However, a recent conflict between Meta and the Oversight Board over content moderation in the Ukraine war raised questions about the organization's independence. Meta had requested a policy advisory opinion (PAO) from the Board on the application of the content moderation system during wartime and subsequently revoked the request. According to media vehicles, the period coincides with the Russian government's declaration that Meta was carrying out extremist activities, which occurred shortly after the platform temporarily allowed calls for violence against Russian soldiers.[8] The tension could be seen as suggestive of the platform's decisions on sensitive issues and could weaken the Board's autonomy, which requires a total freedom untied of the company to make a decision, without fear of reprisals.

Lack of independence to decide on speech and human rights distorts the premise of accountability and due process based on impartiality by offering procedural rights to users

[5] See How Facebook undercut the Oversight Board. The Verge. Casey Newton. 12.05.2022. Available in: https://www.theverge.com/23068243/facebook-meta-oversight-board-putin-russia-ukraine-decision

[6] See How Facebook undercut the Oversight Board. The Verge. Casey Newton. 12.05.2022. Available in: https://www.theverge.com/23068243/facebook-meta-oversight-board-putin-russia-ukraine-decision

[7] See Barrie S, 'Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights Based Approach to Content Moderation' (2020) 43 939. Available in: https://carrcenter.hks.harvard.edu/files/cchr/files/facebook_oversight_board.pdf

[8] See How Facebook undercut the Oversight Board. The Verge. Casey Newton. 12.05.2022. Available in: https://www.theverge.com/23068243/facebook-meta-oversight-board-putin-russia-ukraine-decision

when in truth interfering with their rights. This scenario leads to the question: who watches the watchers? Who watches Oversight Board?

b) Limitation of the scope of the oversight

Platforms have gradually positioned content moderation in a central role of both troublemaker and problem solver, which has raised demands on accountability of content moderation. And this has made the Oversight Board become one of the most important - or at least most visible - battlefield for free speech, human rights in general, and the promotion of democratic content in the context of platforms, even though most of these issues that Meta faces are not fixed by the Board.

While content moderation is a key piece of protecting free speech and human rights online, other issues are equally or maybe more relevant to address main challenges of the digital public sphere, such as competition, and the extraction of personal data to the microtargeting of political adds, to the promotion of some sort of content to the detriment of others and to the recommendation of content. These strategies used by companies to increase the engagement of users[9] also plays a crucial role in the free flow of information and in the virality of malicious content online and, thus, requires accountability, through access to data by journalists and media, for instance[10] - in order to receive mitigation responses.

Unfortunately, these other issues were not included in Meta's project to being hold accountable in their online products and services. It is worth noting that the limited scope of the oversight body has framed many of the global discussions on platforms accountability in terms of content moderation accountability, which contributes to narrow the meaning of platform accountability. The reflection of this semantic meaning could be observed partly, for instance, in the process of UNESCO's Guidelines for Regulating Digital Platforms.[11]

[9] Unver, H. Akin. "Digital challenges to democracy: Politics of automation, attention, and engagement." Journal of International Affairs 71.1 (2017): 127-146.

[10] Bastos M, Mercea D. 2018. The public accountability of social platforms: lessons from a study on bots and trolls in the Brexit campaign. Phil. Trans. R. Soc. A376:20180003.http://dx.doi.org/10.1098/rsta.2018.0003

[11] See Guidelines for regulating digital platforms: a multistakeholder approach to safeguarding freedom of expression and access to information. UNESCO. February 2023. Available in: https://unesdoc.unesco.org/ark:/48223/pf0000384031

c) Effectiveness of the body to promote platform accountability

Although it provides a structure of participation through the public comments it receives and through reasoned decisions, Oversight Board seems to lack effectiveness in providing accountability about Meta's content moderation system. Since the Oversight Board evaluates individual paradigm cases of potential errors in the company's decisions - something similar to judicial review - the body has not the ability to identify systemic problems in content moderation[12] and to disclose information about it. Thus, the Oversight Board could tackle the consequences, but fails to tackle the roots of the problem. As Evelyn Douek suggests, focusing on individual review leads us to miss the "standard picture" of the content moderation process.[13] If the Board is not able to identify the deficits of the system and make it transparent, then the process may not be relevant and effective to address the harms that justify claims for accountability.

The point this section of the paper pinpoints is that the information provided by the review of the Oversight Board does not offer tools to understand the inconsistencies of Meta's operations and the legitimacy and legality of its influence in communications, which impairs meaningful intervention.

4. ALTERNATIVE APPROACH

The question is whether there is a qualified alternative to the current Oversight Board that can counter the criticisms?

- When considering possible alternatives to the Oversight Board, it is first necessary to consider whether a global solution should be pursued or
- whether nationally specific solutions should be sought or
- we can improve the existing mechanism of the Oversight Board to make its decisions more informed and fairer.

Setting up country-specific boards could be considered, which would in turn deal with the concerns of their respective region. In this way, precise regulations could be created at the national level, which would give such smaller bodies stricter guidelines, but also give them more room for maneuver in terms of implementation. By creating several smaller bodies that act on a country-specific basis, the individual cases can be dealt with in a more intensive and exhaustive manner. There would not be one large body responsible for many different cases

[12] Evelyn Douek, 'Content Moderation as Systems Thinking' (2022) 136 Harvard Law Review 526.

[13] ibid.

with national overlaps. The small bodies could decide on cases in advance on the basis of state-regulated guidelines and then hand them over to a higher, globally overarching body for final review. On the one hand, this would have the advantage that the cases would pass through several instances and the review would be based on several experts. Furthermore, the specific regulations of different states could influence the global jurisprudence of content moderation with respect to advancing human rights.

A few other considerations to improve the existing mechanisms of the decision-making process of the oversight board:

- Expanding the existing oversight board in relation to the problem of independence. This suggestion is hinting at creating more layers of adjudication within the oversight board to ensure transparency and fair process.
- Another suggestion from our side is to have involvement of various stakeholders in the decision-making process. Representatives of other stakeholder groups such as civil society organizations, advocacy groups, academicians/ experts, could be deployed in the decision-making process of the Oversight Board. This would allow us to appreciate the issue at hand at the Oversight Board from a comprehensive point of view and not in silos.[14]
- Another suggestion from our side would be to have a system within the Oversight Board where a panel of academicians is being appointed and on every novel issue of content moderation, they submit a report. And the board must mandatorily consider the report while rendering the decision.
- The underlying issue with regulating content moderation is of ex-post error correction and we propose that a separate body should be constituted of academicians/experts within the Oversight Board. This body would be entrusted with the task to formulate substantive core requirements which would leverage private self-regulation of platforms and transform them into public regulatory spheres.

[14] David Morar& Bruna Martins dos Santos, The Push for Content Moderation Legislation Around the World, BROOKINGS (Sept. 21, 2020), https://www.brookings.edu/blog/techtank/2020/09/21/the-push-for-content-moderationlegislation-around-the-world/

5. FINAL CONSIDERATIONS

The Oversight Board within such a private self-regulation of freedom of expression in the user-platform-operator relationship, should not be established as a model of a parallel system to judicial legal protection, since it is not entitled of some judicial guarantees, such as independence.[15] The Oversight Board has been understood as a valid and legitimate body to provide accountability of Meta's decisions on content and this has jeopardized claims for accountability of platforms in a more broader and deep sense, which encompasses their business model based on data extraction and user engagement. In this sense, it is relevant to consider alternative approaches to make platforms more responsible and transparent.

We believe that the Oversight Board has made remarkable contributions to respect and uphold human rights. With its continual interpretative guidance provided to Meta, it will eventually shape private companies content moderation policies around advancing human rights. The board also has the potential to collaborate with regional lawmakers in regulating online platforms significantly with the respect to due-diligence requirements. This informed engagement of Meta with the Board would gradually impact the private online tech players and substantially help them to adopt and practice the values of transparency, accountability, universal access and sharing and fair process.[16]

As the Oversight Board is bound to have long lasting impact in shaping the jurisprudence of advancing online human rights, it becomes imperative that there is deep engagement with other stakeholders such as civil society, advocacy groups and academicians and researchers. The decisions rendered by the Oversight Board would have the greatest impact if all such stakeholders could be deployed in decision making, which will substantially persuade states and the private tech players to bring about meaningful change in society.

[15] Dr. Brosch; Alles neu macht das Meta Oversight Board. MMR 2021.

[16] See, e.g., Molly K. Land & Laurence R. Helfer, Value Pluralism and Human Rights in Content Moderation, LAWFARE (Oct. 27, 2022), https://www.lawfareblog.com/valuepluralism-and-human-rights-content-moderation

WHY IS TRANSPARENCY REPORTING BY ONLINE PLATFORMS IMPORTANT TO THE ACCOUNTABILITY OF PLATFORMS AND SECURING THE RIGHTS OF INTERNET USERS?

by

Bernardo Accioli, Ishita Tulsyan and JosseAmanieu

1. INTRODUCTION

One can only judge a ruling by its reasoning. While the term transparency can accommodate a wide range of aspects, in the context of platform governance for the purpose of this paper, it is mainly delimited to two: **(*i*)** direct platform-to-user/public relaying of information on questions of content moderation, data collection, algorithmic processes, etc.; and **(*ii*)** periodic transparency reporting from the platform to the public and/or the government.

Transparency as an ideal becomes especially valuable for online platforms in today's times where being online is not a matter of option anymore,[1] but a must in a massively digitised society. Its absence harms the rights of the users, and, to tackle this, effective policy formulation is the need of the hour. As social media completely integrates the user's *entourage*, mimicking the physical world, certain questions arise: if conviction or imprisonment requires following due process to safeguard individual rights in the offline world, why do automated decisions get to temporarily remove or permanently ban users online without adequate safeguards?

This paper thereby attempts to approach transparency regarding account retention and suppression decisions made by the platforms. It attempts to highlight the current legal frameworks of India, Germany and Brazil in this area. This issue will be addressed from the perspective of a hypothetical case and how the three legal systems approach the matter.

In the end, we make recommendations about an ideal platform policy. Specifically, we recommend that countries institute a requirement for periodic transparency reports from platforms. We also recommend that countries establish an independent institution, akin to Germany's FSM, to transparently review content blocking.

[1] Rekha Pathak, "The role and functions of social media in socialization" (ResearchGate December 2019) <337707053_The_role_and_functions_of_social_media_in_socialization> accessed 17 March 2023.

2. CURRENT LEGAL APPROACH

2.1 INDIA

In India, the main law to govern the transparency requirements on platforms is the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 (hereinafter, the Rules).[2]

Theoretically, if the hypothetical of a content moderation event were set in India, as a significant social media intermediary (hereinafter, SSMI), the platform would be mandated to provide the complainant with a constant track of their complaint and a reason for the action or a lack of it.[3] Similarly, the originator of the content would also be provided with the reasoning for the action being taken prior to the actual action.[4] Further, the recent amendments to the IT Rules[5], provide for a central government-appointed Grievance Appellate Committee for user appeal against platform decisions on content moderation.

While this is the per law analysis, content moderation reports during the first round of report release have shown a lack of compliance and inadequate reporting by platforms.[6] Further, the governance mechanism under these rules has largely been deemed undemocratic.[7] Even with this partial reporting, a fact that has been made clear is that government surveillance is a constant reality in suggesting content to be removed.[8] Thus, with the high degree of discretion available to the government, the Rules may be easily manipulated to suit the whims of the government, especially regarding the threshold of SSMIs, which directly influences the transparency and other measures they must mandatorily undertake.

Further, there have been long-standing cases where the government itself refused to provide the reasoning for blocking orders under section 69(A) of the IT Act.[9] However, with the

[2] Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021.

[3] IT Rules 2021, s 4(6).

[4] IT Rules 2021, s 4(8)(a).

[5] Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2022.

[6] "Big Tech releases its transparency reports in compliance with the IT Rules: Here's what we found" (Internet Freedom Foundation, 16 July 2021) <https://internetfreedom.in/big-tech-releases-its-transparency-reports-in-compliance-with-the-it-rules-heres-what-we-found/> accessed 17 March 2023.

[7] "Deep dive : How the intermediaries rules are anti-democratic and unconstitutional" (Internet Freedom Foundation, 27 February 2021) <https://internetfreedom.in/intermediaries-rules-2021/> accessed 17 March 2023.

[8] Ibid.

[9] Tanul Thakur v UOI; "Delhi High Court issues notice in the blocking case of satirical website" (Internet Freedom Foundation, 25 January 2023) <https://internetfreedom.in/dhc-issues-notice-in-website-blocking-case/> accessed 17 March 2023.

advent of the Digital India Act, which seeks to revamp the IT Act, there will certainly be changes witnessed that one must wait for in the policy domain.[10]

2.2 GERMANY

Germany addressed issues of transparency by implementing several laws and policies to promote platform transparency. In the present case, the Netzwerkdurchsetzungsgesetz (NetzDG) would be the main relevant law. The NetzDG codifies online platform regulation and transparency, particularly in terms of content moderation and removal.[11]

It is already disputed whether the German legislator is even authorised to impose such regulations on platforms based abroad, as this could violate the European "country of origin principle" which is stipulated in the e-commerce directive of the European Union.[12] Platforms such as YouTube deny applicability but still implement the regulations to avoid potential legal consequences. However, in 2022, a German Administrative Court declared central provisions of the NetzDG to be contrary to European law.[13]

If, nevertheless, the applicability in the present case is assumed, it is still necessary to distinguish whether the requirements of the law are met. For Instance, the NetzDG is only applicable to social media platforms that have at least two million registered users.[14] These platforms are required to remove or block illegal content within 24 hours when receiving a complaint, or within 7 days for more complex cases.[15]

In addition to that, social media companies are required to inform users whose content has been removed or blocked of the reasons for such action and about the outcome of the review

[10]<https://www.meity.gov.in/writereaddata/files/DIA_Presentation%2009.03.2023%20Final.pdf> accessed 17 March 2023.

[11] Enforced October 2017 <https://www.bmj.de/DE/Themen/FokusThemen/NetzDG/NetzDG_EN_node.html> accessed 17 March 2023.

[12] This principle suggests that a company or service provider should comply with the regulations and laws of the country where it is established, rather than the regulations and laws of the countries where its services are being consumed or accessed; Marc Liesching, "Gilt das NetzDG für Facebook, Youtube und Twitter?" <https://community.beck.de/2020/02/11/gilt-das-netzdg-fuer-facebook-youtube-und-twitter> accessed 17 March 2023.

[13]The background to the lawsuit against the Federal Republic of Germany was a new provision in the NetzDG that provides for the transfer of user data such as IP addresses or port numbers to the Federal Criminal Police Office if illegal content has been removed or blocked; Friedhelm Greis, "NetzDGistteilweiseeuroparechtswidrig" (1 March 2022) <https://www.golem.de/news/gerichtsurteil-netzdg-ist-teilweise-europarechtswidrig-2203-163530.html> accessed 17 March 2023.

[14]NetzDG 1(2); The NetzDG applies to a range of illegal content, including hate speech, defamation, and incitement to violence.

[15]NetzDG 3(2) Nr. 1,2,3; NetzDG 3a(3).

process, including whether any content was removed or blocked as a result of the complaint.[16]

The law also provides a mechanism for users to challenge the removal or blocking of their content, including in cases where the removal was due to false accusations.[17] The platform is then required to review the request and make a determination within 7 days.[18] In these 7 days, a so-called "regulated self-regulation" facility can also help to find an appropriate decision, especially in difficult cases.[19]The first and currently only facility of this kind is the "FSM".[20]

The NetzDG has been criticised by some for potentially infringing on freedom of speech and putting too much responsibility on social media companies to police content. Furthermore, it is stated that it does not have any real practical relevance, as most of the content is already deleted by the platforms for violating their terms of service. However, supporters argue that this legal framework is necessary to combat illegal content online and to protect individuals from harmful content.[21]

[16]NetzDG 3(2) Nr. 5; This means that if an Influencer's account was suppressed, the platform would need to provide a specific explanation for the action taken, including the reasons for any complaints or reports that led to the suppression.

[17]NetzDG 3b among other paragraphs.

[18] If the company determines that the content was removed or blocked in error, it must be restored. If the social media company fails to restore the content or fails to respond to the user's request within these 7 days, the user can bring the matter to the attention of the German Federal Office of Justice However this was one aspect that was found to be in violation of European law by the German Administrative Court.

[19] The goal of these facilities is to promote industry responsibility while also ensuring that the public interests are protected. Regulated self-regulation is often used in industries where the government recognizes that self-regulation is an effective and efficient way to ensure compliance with regulations, and where there is a high level of expertise and knowledge within the industry itself.

[20] The "FreiwilligeSelbstkontrolle Multimedia-Diensteanbieter (short FSM, english: Voluntary Self-Regulation of Multimedia-Service-Providers) was established in 2020 and includes a panel of experts who can decide on difficult deletion cases. The decisions are transparently published on their website: <https://www.fsm.de/fsm/netzdg/> accessed 17 March 2023.

[21] At the European level, regulation is also considered essential, and a European regulation is imminent through the Digital Services Act (DSA) of the European Union (scheduled to take effect on 1 January, 2024). The DSA is also the reason for the growing uncertainty about the future applicability of the NetzDG in its current form. Unlike the NetzDG, which applies only to social networks, the DSA will regulate all digital services.

2.3 BRAZIL

In Brazil, the subject of this article would not have an obvious solution directly arising from the legal text.[22] But it is possible to say that most likely, if the user reached out to court, he/she would be able to get her deleted account back, in the event the company were unable to accurately demonstrate her violation. In order to not be held accountable, the platform would as well have to prove that it gave the user adequate means to defend his/herself.

As none of the mentioned laws address the issue directly, the obligation for the platform to expose the reasons for removing the content or the profile, and to provide due legal process, comes from an interpretation of the law by case law. The courts have,[23] therefore, found that platforms must provide some sort of due process of law for account removal, specifically presenting the reasons for the takedown. The rulings are usually split between the interpretation of the Brazilian Civil Rights Framework for the Internet (MCI),[24] and the interpretation of the Code of Consumer Defence and Protection,[25] or, at most, the Brazilian Civil Code, based on the general clause of contractual good faith.[26]

At present, there is a bill (Projeto de Lei, "PL" 2630/2020) that addresses "Freedom, Responsibility and Transparency on the Internet".[27] The current version of the Bill stipulates the requirements of the content moderation procedure, such as notification and the possibility of opposition or appeal. If the content is wrongly classified as infringing, the subsequent decision acknowledging the platform's mistake should be publicised. By this bill, platforms would have to specify how their content moderating algorithm works and how it flags

[22] In Brazil, the directly applicable law is not very clear, as there are laws in force (i) that guarantee net neutrality and the accountability of platforms for third-party generated content (similar to the US Section 230), the Marco Civil da Internet (Brazilian Civil Rights Framework for the Internet, Law n. 12. 965/14); (ii) that ensure personal data protection (Lei Geral de Proteção de Dados, Law n. 13.709/2018); (iii) that protect the consumer (Code of Consumer Defence and Protection, Law n. 8.078/90); and (iv) that provide for the incidence of good faith and contractual loyalty in civil matters (Código Civil, Law n. 10.406/02). The Brazilian Constitution also sets forth due process of law (art. 5, LIV), as well as the direct applicability of this fundamental right, in art. 5, §1, but normally these issues are not directly addressed by court rulings on the matter.

[23] Please note that Brazil is a country of Roman-Germanic tradition, and therefore case law, even that of higher courts, does not have, as a rule, binding effect. There are some hypotheses in which a judgement will be granted binding effect, but this is an exceptional scenario.

[24] *Twitter Brasil v Vareta* [20022], Ch 17 TJSP

[25] *Facebook v Ferreira* [2022] Ch 34 TJSP

[26] In a concrete case, the Rio de Janeiro State Higher Court ruled that the banning of a driver from the Uber Platform violated good faith (art. 422, Brazilian Civil Code) because no justification was given and the driver's right of defence was not respected. *Uber do Brasil v Fernandes* [20223] Ch 16 TJRJ

[27] Chamber of Deputies of Brazil, Projeto de Lei, Projeto de Lei 2630/2020. <https://www.camara.leg.br/propostas-legislativas/2256735>. Accessed 17 March 2023

harmful content. Stakeholders have classified this provision as concerning, as it may "*hand the gold to the bad guy*",[28] because once in possession of this information, disinformation agents can reverse engineer the algorithm and learn how to escape the platforms' filter. Transparency can be, then, a double-edged sword.

The bill also states that social network providers must produce six-monthly[29] transparency reports, based on provided parameters, and subjected to revision by an "Internet Steering Committee". They would inform about active intervention in accounts and third-party generated content which may entail deletion, unavailability, reduction of scope, flagging of content and other types of content that restrict freedom of expression.

Given the presidential elections of 2022, PL 2630 seems to have been put aside at first. Also, the current Minister of Justice is prone to rewriting the PL from scratch, as his main focus is not transparency, but rather addressing disinformation and democracy. Despite the Minister's onslaughts, PL 2630 is still being advocated by the Social Communication Secretariat of the Presidency (Secom).[30]

3. POLICY

The existing legal systems all share a common goal: the effective protection of user data and users' rights. The challenge lies in finding the most effective ways to establish transparency options that balance the independence of platforms, protection of user rights, and the state's obligation to provide a legal framework. A comparative legal analysis can provide insights into the most effective solutions and allow for their examination.

One possible efficient solution would be to establish independent institutions, similar to the German-regulated self-regulation body (known as FSM). Such institutions can minimise state intervention and represent user interests through an independent panel of experts, who can handle difficult individual cases and provide a human element in deciding whether illegal content is present. This approach aims to ensure an independent review of content blocking, which is sometimes just carried out by algorithms.

---

[28] Rio de Janeiro Institute for Internet and Society, *9 pontos de atençãosobre o PL das Fake News (PL 2630/20)* (White Paper, 2022) <https://itsrio.org/wp-content/uploads/2022/04/9-pontos-de-aten%C3%A7%C3%A3o-sobre-o-PL-das-Fake-News-PL-2630_20.pdf> Accessed 14 March 2023

[29] The original draft envisaged quarterly reporting.

[30] Guilherme Caetano, 'Governo Lula se divide sobre PL das Fake News entre cobranças de punição e transparência a plataformas: Ministério da Justiça temvisãomaispunitivista, enquantofoco da Secom é sobrealgoritmosusadospor redes sociais. executivobuscaconsenso para enviarsugestõesao relator' *O Globo* (Rio de Janeiro, 12 March 2023)

Another important policy aspect would be to implement periodic transparency reports. Content moderation decisions made by online platforms may impact users' rights to freedom of expression and privacy. In many cases, users are not notified of the reasons for the removal of their content or the suppression of their accounts, which can lead to unfair outcomes. Transparency reports would help to address these issues by providing information on the criteria used for the content moderation. Also, they can aid informed policy-making to curb other online harms, such as hate speech and disinformation. At the same time, the reports can provide policymakers with data on the prevalence and nature of these harms, as well as the effectiveness of current policies and measures in addressing them.

These approaches serve as a compelling demonstration to platforms and policymakers that transparency can yield significant benefits for all stakeholders. Hence, it is highly desirable to endeavour towards establishing such possibilities in a global context.

4. CONCLUSION

Transparency reporting is crucial for ensuring accountability of platforms and securing the rights of Internet users. Transparency reports and independent institutions seem the most effective. In order to fully exploit the possibilities of transparency, the communication between legislators, platforms and not to forget users must be expanded and promoted. Policymaking is not only about defending and enforcing rules – it is also about creating new possibilities. The purpose of transparency is not for politics and platforms to act against each other. Both – legislators and platforms – should be able to benefit from the transparency options in the best possible way and thus lead to an understanding and protection of the users' rights.

SHOULD ONLINE PLATFORMS BE REQUIRED TO PROACTIVELY MONITOR FOR UNLAWFUL CONTENT?

by

Aastha Singh, Lukas Kellermeier, and Natalia Gigante

1. INTRODUCTION

Proactive monitoring means a business is continuously searching for potentially harmful or unlawful content. Proactive monitoring is done to ensure that material available on the platform is legal. Content such as child sexual abuse material and terrorism content and copyright infringement content is often subject to proactive monitoring. Such monitoring can be done in order to make sure that harm does not happen. It is more precautionary than reactive.

This piece discusses the trends around the world and the advantages along with consequences of a proactive monitoring regime. In our final suggestions, we explore the possibility of a graded content-based system for proactive monitoring.

2. CURRENT STATUS OF MONITORING

Currently there are different practices of identifying prohibited content. Most deployed filters are designed to find (even cropped) duplicates of known, specific content such as images, audio, or videos (e.g., PhotoDNA for child sexual abuse images or videos, and to find violent extremist images or videos).

Duplicate-detection filters for written text are technically even simpler and have existed for decades – their basic function is familiar to anyone who has used the ‘find’ function in a browser or a text editor like Microsoft Word. Another filter practice is threat profiling based not on content but on uploaders’ behaviour (using spam fighting tools to flag suspicious patterns of contacts, followers, or posting locations).

Proactive Monitoring challenges the ‘notice and takedown procedure’ and the safe harbour paradigm of intermediary immunity. The assumption of the paradigm is based on the fact that platforms are “just” intermediaries and have no control over the material sent

on their networks. Intermediaries remove content upon receiving 'actual knowledge' in the form of court order or user complaint.[1]

In the proactive monitoring regime, intermediaries may be required to independently identify and remove content even before a complaint by user or court order. This could lead to:

- platforms gaining discretionary powers
- platforms may err on the side of caution and remove content that has a remote risk of illegality
- become a hindrance for free speech.

Proactive Monitoring can also undermine privacy. Requiring a platform to continuously monitor all users could easily make it a surveillance tool.

STAKEHOLDERS AND FUNDAMENTAL RIGHTS (LEGAL INTERESTS) AFFECTED

As explained by Jack Balkin, the 21st century model of freedom of speech can be explained as a triangle formed by nations, the internet companies, and users.[2] According to him, this configuration raises new issues, in relation to the moderation of content, which may result in censorship, absence of transparency and lack of opportunities for the users to appeal against removal decisions.[3] Additionally, the users may be in a vulnerable position and not have sufficient remedies to protect their rights.[4]

At this point, it must be noted that fundamental rights do not (directly) bind private entities such as platforms. Rather, they can have an indirect third-party effect in certain situations.Nonetheless, faulty, or inaccurate filtering processes may (through platform behaviour) threaten fundamental rights such as:

- freedom of expression and information (where users' content is incorrectly restricted)

[1] Art. 12-14 e-Commerce Directive; Section 79(3) India IT Act; Art. 19 Marco Civil.

[2] Jack M Balkin, 'Free Speech Is a Triangle' (2018) 118 Columbia Law Review 2011.

[3] ibid.

[4] ibid.

- privacy and data protection (e.g. filters might require scans of innumerable other people's pictures; in general, proactive monitoring may infringe on privacy by requiring platforms to monitor all user accounts constantly for unlawful or suspicious activity, effectively requiring platforms to engage in a form of mass surveillance)

- rights to a fair trial and effective remedy (for people whose online expression and participation are 'adjudicated' as violations and terminated by platforms - this could be a violation of the law if the removal of lawful content leads to damages for the user who posted it)

- rights to equality and non-discrimination before the law (two recent studies, for example, found that when automated content filters attempt to parse human language, they disproportionately silence lawful expression by members of minority or marginalized racial and linguistic groups[5])

3. CONTENT-AREA DETERMINED PROACTIVE MONITORING

To act against unlawful content, it is important to highlight that different issues demand different approaches. For example, content such as child sexual abuse material is easier to identify and also more likely to cause damages in society. In these cases, the error rates associated with incorrect removals may be lower and platforms are more likely to proactive monitor and remove the contents. Other violations, such as copyright, industrial property infringements, such as the sale of counterfeit products and misinformation, demand a deeper analysis.

There's also the possibility to monitor, but not to remove the content. For example, to protect copyright, Google (YouTube) uses proactive monitoring to identify allegedly infringing content but then, instead of always removing the content, it notifies copyright owners and allows them to present take-down requests against infringing content. In this sense, YouTube utilises proactive monitoring technology but the user chooses if they will pursue the matter towards removal or not.

[5] Mixed Messages?, 'The limits of automated social media content analysis' (Center for Democracy and Technology (cdt), November 2017), <https://cdt.org/files/2017/11/Mixed-Messages-Paper.pdf>; Maarten Sap and others, 'The Risk of Racial Bias in Hate Speech Detection' (2019) University of Washington <https://homes.cs.washington.edu/~msap/pdfs/sap2019risk.pdf>.

4. REGULATORY APPROACHES AROUND THE WORLD

INDIA

In 2021, the Indian parliament introduced the Intermediary Guidelines and Digital Media

Ethics Code (2021 IT Rules). Rule 4(4) encourages Significant Social Media Intermediaries [SSMIs] to use "technology based measures" to proactively monitor content which had been previously found to be in violation of the said rules[6]. Though it also encourages SSMIs to monitor and remove material which should never reach platforms such as child sexual abuse material,[7] it essentially compels companies to adopt automated tools, which monitor the users and their posts across a wide range of subject areas. This leads to error prone filters that restrict lawful online expression. The Rules also put the onus on intermediaries and platform to inform the users of the rules and ensure that they do not "host, display, upload, modify, publish, transmit, store, update or share"[8]any of the restricted types of contents'.

BRAZIL

In Brazil, there is no current legislation obliging the online platforms to proactively monitor for unlawful content. However, this does not mean that there is no liability involved in the activities provided by such platforms.

Before the promulgation of the Brazilian Civil Rights Framework of Internet (Marco Civil), polemic Court decisions were issued concerning this matter, demanding that the platforms remove the contents involved in lawsuits in a generic way, as seen in *Dafra vs. Google*, when the judge considered the online platforms as "*untameable monsters*" and required the platforms to take measures to remove the infringing contents on frequent basis, which demands a practice of active monitoring for specific content.[5]

[6] Rule 3, IT(Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021

[7] Rule 4, Sub rule 4 IT(Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021

[8]Frosio, Giancarlo, The Death of 'No Monitoring Obligations': A Story of Untameable Monsters (June 5, 2017). 8(3) Journal of Intellectual Property, Information Technology and E-Commerce Law (JIPITEC) 212 (2017. p. 200, Available at SSRN: https://ssrn.com/abstract=2980786

In 2014, with the promulgation of the Brazilian Civil Rights Framework of Internet, the liability of the platforms started being determined based on the service provided by the company while paying attention to its capacity to manage the content published on its network.[9] In this sense, if the main service or purpose of the online platforms is to distribute unlawful content, it may be liable.

On the other hand, with respect to contents posted by third parties, online providers would be liable only if they disobey Court orders asking for the removal of content. The exception is when the content is related to private videos involving nudity or sexuality.[10] In this case, the providers should remove the video after a notice and take-down request.

Another exception involves matters involving copyright issues. Marco Civil establishes that the rules applied to the removal of these contents should be handled by a specific law, which was not promulgated yet. In this sense, the liability of the internet providers is determined in the same terms established for all matters, excluding personal videos involving nudity or sexuality.

To avoid the presentation of general requests, which may demand an active monitoring, Marco Civil establishes that the user should indicate the URLs that should be removed. There is no local rule forbidding online platforms, to develop their own policies regarding moderation of content, so they can. Considering the latest political events involving disinformation and democracy, the Brazilian authorities are discussing possibilities of changing the liability rules in Brazil in the future.[11]

[9] TEFFÉ, C. S.; SOUZA, C. A. Responsabilidade civil de provedores na rede: análise da aplicação do Marco Civil da Internet pelo Superior Tribunal de Justiça. Revista IBERC, v. 1, n. 1, p. 9, May 22, 2019.

[10] BRAZIL. Law no. 12965, from April 23rd, 2014.The Brazilian Civil Rights Framework of Internet. Available at. http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.html.

[11] ORTELLADO, Pablo. O fim do artigo 19 do Marco Civil da Internet. O Globo. March 14, 2023. Available at https://oglobo.globo.com/opiniao/pablo-ortellado/coluna/2023/03/o-fim-do-artigo-19-do-marco-civil-da-internet.ghtml

EUROPEAN UNION

Generally, platforms are not liable for third party illegal content. Rather, the so-called notice & takedown system establishes (limited) liability if content is not removed (in a timely manner) despite notification of the platform.

There are numerous regulatory approaches to platform regulation in the EU (DSA, DMA, P2B regulation, DGA, DA, AI Act, DSM Directive), some of which entail proactive monitoring. For example, the Copyright Directive created a de facto filtering mandate by requiring platforms to 'prevent further uploads' of specific works.[12] That Directive provides that decisions to disable access to or remove uploaded content shall be subject to human review.'[13] Furthermore, the drafts of the Terrorist Content Regulation say that hosts using 'automated tools' to assess user content 'shall provide effective and appropriate safeguards' against improper removals, consisting 'in particular, of human oversight and verifications' of filters' decisions – though only in the Parliament draft is such human review clearly mandatory.[14]

In the context of proactive monitoring, Article 15 of the eCommerce Directive (cf. Article 7 DSA) is particularly important. This article states that 'general' monitoring obligations on platforms are prohibited. This leads to the legal question: Where are the boundaries between prohibited 'general' monitoring and permissible 'specific' monitoring? The Austrian Supreme Court asked the CJEU whether orders to block 'identical' or 'equivalent' content were permissible under Art. 15 of the eCommerce Directive.[15] That referral did not ask the CJEU about fundamental rights. The Advocate General (AG) advised that such orders were permissible under certain circumstances. Broadly, the Court held that injunctions requiring platforms to proactively remove both identical and

[12] Article 17.6, Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (so called Copyright Directive) <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019L0790>.

[13] Article 17.9 Copyright Directive, see footnote above.

[14] Article 9.2 Commission, 'Proposal for a Regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online' COM(2018) 640; cf. *Keller*, Facebook Filters, Fundamental Rights, and the CJEU's Glawischnig-Piesczek Ruling(GRUR Int. 2020, 616, 620 with further proof.

[15] Daphne Keller, 'Facebook Filters, Fundamental Rights, and the CJEU's Glawischnig-Piesczek Ruling' (2020) 69 GRUR International 616.

equivalent content are permitted by the eCommerce Directive.[16] Courts could nonetheless issue more specific injunctions to block particular content identified by the court. The CJEU approved an injunction that required internet access providers to block particular websites but did not specify the measures which that access provider must take.[17]

5. THE FUTURE OF PROACTIVE MONITORING: OUR PROPOSALS

To deal with false detections, there must be a procedure for preventing or correcting filtering errors. One solution involves creating a whitelist of known false positives. Any word appearing on the whitelist can be ignored by the filter, even though it contains text that would otherwise not be allowed.

Means of correcting errors might include notifying affected users and allowing them to challenge removals using platform-operated appeal or 'counter-notice' systems. The efficacy of those systems, though, is questionable – and at best they provide a remedy for speakers, but not for users unknowingly deprived of access to information. To prevent this, one could think of a mechanism that informs any user trying to access content that it was pursuant to proactive monitoring that the content had been taken down, enabling the user to challenge the takedown. To protect against errors, 'the national procedural rules must provide a possibility for internet users to assert their rights before the court once the implementing measures taken by the internet service provider are known.'[18] This means that users need to be able to protect their rights (in court).

Another discussed solution is the requirement of human judgment as an element of proactive monitoring regimes. However, this leads to follow-up problems. Even if courts could require platforms to carry out human review, it is unclear how well such review would correct for filters' mistakes. Humans may merely rubber-stamp decisions produced by filters – and have incentives to do so to avoid legal risk. Indeed, researchers have

[16] Case C-18/18 Eva Glawischnig-Piesczek v Facebook Ireland Limited EU:C:2019:821, Opinion of AG Szpunar.

[17] C-314/12 UPC Telekabel Wien GmbH v Constantin Film Verleih GmbH and WegaFilmproduktionsgesellschaftmbH EU:C:2014:192 = GRUR Int 2014, 469, para 64.

[18] C-314/12 UPC Telekabel Wien GmbH v Constantin Film Verleih GmbH and WegaFilmproduktionsgesellschaftmbH EU:C:2014:192 = GRUR Int 2014, 469, para 57.

identified high rates of over removal and error even in purely human-operated notice and takedown systems.[19]

The moderation of content should be very transparent, and the users should be aware of what content is allowed to be posted and the rules applied by each provider. Besides the disclosure of the rules applied, the users also should have been granted with the possibility of presenting appeals against decisions of the removal of content. Both Manila Principles and Santa Clara Principles, developed by civil society organizations, provide a good start in settling rules about moderation, giving attention to the protection of civil rights and the sovereignty of each country. Observing these principles, especially involving transparency, might guarantee a better strategy of monitoring and avoid excessive moderation or over removal.

6. OUTLOOK: CONTENT BASED LEVEL SYSTEM FOR PROACTIVE MONITORING

There are ways to combat illegal content other than requiring upload filters, such as tax-funded government institutions for rapid review of reported content. Following the EU liability rules on artificial intelligence, a tier system could be conceived with regard tocontent filters. The corresponding EU regulation distinguishes between a total of four risk classes (unacceptable, high, low, and minimal risk).[20]

With respect to the regulation of content filters, a differentiation according to various categories could be developed: the filtering of content that cannot be legal under any circumstances (such as child pornography) is probably quite unproblematic - in this case, it should only be ensured that the recognition systems (primarily image recognition) are as technically mature as possible.

[19] Cf. *Keller*, Facebook Filters, Fundamental Rights, and the CJEU's Glawischnig-PieszekRuling(GRUR Int. 2020, 616, 622 f. with further proof.

[20] AI practices that are considered unacceptable, for example because they violate fundamental values of the EU, will be prohibited (Art. 5 AI Regulation-E). For AI systems with a high risk, minimum requirements apply (Art. 8 et seq. AI-Reg-E), which must be fulfilled by providers and users of the systems (Art. 16 et seq. AI-Reg-E). In addition, irrespective of the risk class, transparency requirements apply to certain AI systems that exhibit specific risks of manipulation (Art. 52 AI Regulation-E). In contrast, AI systems with a low or minimal risk are not subject to regulation. However, providers of such systems may voluntarily adhere to codes of conduct (Art. 69 AI Regulation-E).

Another category could be content that is so sensitive, if it is indeed illegal, that it should first be blocked by default, followed by human review as quickly as possible (for example, terrorism videos that appear to contain executions). This may be different for less sensitive content (such as possible insults or copyright infringements), which is usually more ambivalent and therefore carries a higher risk of false evaluation - here, no deletion should take place by default, but a mechanism for verification should be implemented (such as a reporting option to then proceed according to the notice & takedown principle).

It would also be worth considering assigning a special position to certain accounts (such as newspapers) after a registration procedure, so that their content is generally not blocked by default unless they are reported and a review or court decision concludes that they have exceptionally committed a legal violation (for example, due to violation of personal rights or copyright). A similar privilege should then apply to government institutions and their accounts/content.

ARE SPECIAL MEASURES NEEDED TO COUNTER THE HARMS ARISING FROM TECH-FACILITATED GENDER-BASED VIOLENCE (TFGBV)?

by

Fee Zimmermann, Luize Pereira Ribeiro and Navdha Sharma

1. INTRODUCTION

Gender-based violence has been in existence for decades[1]. It too often diminishes and assaults some individuals, mainly due to their gender[2]. This pervasive structure allows victims to face immense obstacles that consequently and severely affect their access to equal opportunity, fair treatment, safety, and fundamental human rights like dignity, health, education, privacy, among other inalienable rights[3].The issue of gender-based violence translates online as well and encompasses a range of abusive behaviours, which arse usually assumed to be limited to non-consensual intimate image abuse colloquially referred to as 'revenge porn.' However, it extends to practices like sexualized photoshopping, domestic violence, stalking, harmful comments, sextortion, voyeurism, and upskirting, among other forms of abuse.

It is the need of the hour to develop comprehensive policies and measures aimed at preventing and addressing online gender-based violence. Online violence against women is often dismissed as it is wrongly considered to be something which does not impact theirreal life. However, time and again it has been noted that online violence often translates to offline violence as well, like the case of a female gym employee in Serbia wherein a man resorted to online threats against her before physically assaulting her[4].

[1]Hrick P, 'The Potential of Centralized and Statutorily Empowered Bodies to Advance a Survivor-Centered Approach to Technology-Facilitated Violence against Women', The Emerald International Handbook of Technology-Facilitated Violence and Abuse (Emerald Publishing Limited, 2021). Available at: https://www.emerald.com/insight/content/doi/10.1108/978-1-83982-848-520211043/full/html

[2] Council of Europe, 'No space for violence against women and girls in the digital world' (2022). Available at: https://www.coe.int/en/web/commissioner/-/no-space-for-violence-against-women-and-girls-in-the-digital-world#

[3] Ibid.

[4] SHARE Monitoring . Available at: https://monitoring.labs.rs/data?caseId=3456) (Accessed 14 mar. 2023)

Considering this environment, the present study aims to explore what specific and impactful measures are urgently called for to counter the profound harms arising from TFGBV. Firstly, the paper attempts to provide a brief overview of the types of harm faced by the victims of TFGBV, the vital needs at stake for vulnerable groups who face this violence under specific and disproportionate consequences and difficulties. Then this paper delves into the technical, legal and enforcement issues related to tackling TFGBV. Thereafter, content moderation, implementing laws, and establishing enforcement bodies is explored as a possible solution to the pervasive problem to TFGBV, while also weighing it against the limit of freedom of expression, along with the debate of overblocking[5]. Lastly, after examining this multifaceted panorama, the study presents a series of possible comprehensive solutions and policy recommendations to counter the harms arising from TFGBV, particularly for vulnerable groups facing multiple axes of oppression.

2. GROUPS AT PARTICULAR RISKS

Although the internet does not itself discriminate based on gender, women undergo far more severe forms of online violence and abuse. To add to this, women belonging to marginalized groups are at a greater risk of experiencing tech-facilitated violence. To better explain the dire situation, in various studies it was found that one in five women in Canada experience some form of online harassment in 2018[6]. In France, 15% women experienced cyber harassment[7]. In the United States, as per the Pew report of 2017, women have been found twice as likely as men to say that they have been targeted due to their gender. In Pakistan, as per the Hamara Internet study's report, 40% of women have faced some sort of harassment.

To narrow it down, women belonging to structurally and historically marginalized groups such as the LGBTQIA+ community, religiously persecuted groups, women with disabilities and women of colour experience far higher rates of online abuse on Twitter.[8] Surprisingly,

[5]Monea A, '3. Overblocking' [2022] The Digital Closet. <https://digitalcloset.mitpress.mit.edu/pub/i2a3g68w/release/1>.

[6] Gender based violence and unwanted sexual behavior in Canada (2018)<https://www150.statcan.gc.ca/n1/ daily-quotidien/191205/dq191205b-eng.html>

[7]European Union, 'Cyber violence and hate speech online against women' (2018)<https://www.europarl.europa.eu/RegData/ etudes/STUD/2018/604979/IPOL_STU(2018)604979_EN.pdf>

[8]Amnesty Internation, 'Toxic twitter - triggers of violence and abuse against women on Twitter' (2022)al.<https://www.amnesty.org/en/latest/news/2018/03/online-violence-against-women-chapter-2-3/>

women of colour are 84%[9] more likely to receive abusive messages. Further, in a recent incident in Iran, women posting pictures of themselves without a hijab were not only subject to online abuse, but were prosecuted by the state and later imprisoned[10]. When considering the intersection of race and gender, it becomes evident that black women are more likely to be the targets of online harassment compared to their white and male counterparts. This phenomenon is highlighted in a study on "Algorithmic misogynoir in content moderation practice" by Brandeis Marshall:

"Shireen Mitchell's 2018 Stop Online Violence Against Black Women report showed how online campaigns using Facebook ads were created to disparage Black girls and women with sexualized memes, hashtags, and fake accounts to help spread disinformation ahead of and during the 2016 U.S. Presidential Election. Content moderation can work towards creating inclusive, welcoming spaces for Black women, but current practices embrace misogyny and then deploy it algorithmically."[11]

As rightly pointed out by Rachel Hatzipanagos,[12] online violence is not confined to the digital realm and often spills over into the real world and victimisesmarginalised groups. Thus, it is imperative that tech companies and states take action against harmful speech on their platforms because the more it is tolerated, the more it becomes normalised. Some effective tools in this fight are content moderation, legal regulations and compliance which can prevent violent behaviour from becoming pervasive.

3. ISSUES PLAGUING CONTENT MODERATION

The current system against TFGBV continues to be ineffective due to three main reasons: technical, legal and enforcement issues.

---

[9] Council of Europe, No space for violence against women and girls in the digital world (2022)<https://www.coe.int/en/web/commissioner/-/no-space-for-violence-against-women-and-girls-in-the-digital-world#>

[10] *'Iran Arrests Eight for "un-Islamic" Instagram Modelling' BBC News (16 May 2016) <https://www.bbc.com/news/world-middle-east-36302405>*

[11] Marshall B, 'Algorithmic Misogynoir in Content Moderation Practice' [2021] Heinrich-Böll-Stiftung European Union

[12] How online hate turns into real-life violence (2018), Rachel Hatzipanagos. Available at:

Hatzipanagos R, 'Perspective | How Online Hate Turns into Real-Life Violence' Washington Post (30 November 2018) <https://www.washingtonpost.com/nation/2018/11/30/how-online-hate-speech-is-fueling-real-life-violence/>

The technical issue is essentially concerned with difficulty in moderating content due to a variety of reasons such as lack of context, cultural differences, language barriers, etc. The issue of language barrier remains prominent as it creates difficulty in filtering the content using "keywords" due to nuances in different languages. As most AI models are originally trained in English and first translate the content from any language to English and then filter via keywords to see if it is likely to be harmful. Thus, due to cultural differences and different contextual meanings according to various languages, the model fails to accurately identify harmful content.

In terms of legal issues, the lack of regulation in TFGBV cases is still incipient as countries' criminal codes severely lag behind in updating the laws according to the increasing shift from offline to online conversations. The lawmakers have failed in criminalising posting of harmful content targeted at specific groups online. It is high time that a law is enacted in nations across the world to assign responsibility to individuals and platforms. For example, India introduced Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 to hold social media intermediaries accountable.

Lastly, even if lawmakers were to implement laws, enforcement issues would remain a huge problem due to the issue of actually holding platforms liable and imposing sanctions. In a globalised world, platforms hold significant power due to a large user base, thus shutting them down due to non-compliance is a far-fetched and unlikely solution. In such a case, the laws need to be structured in a manner in which they ensure platforms' compliance with them while also balancing economic, user and national interests.

4. THE WAY FORWARD

Firstly, it is important that users are informed of the reporting measures and grievance redressal mechanisms in place so that they are aware of their rights and the conditions under which their participation on a particular platform can be limited. Thereafter, as a first line of measure, it is of utmost significance that the existing issues in the AI models are addressed to make them more robust and accurate in identifying harmful content. There are several techniques, which identify, match, predict and classify pieces of content (e.g. text, audio, image or video). One critical aspect to develop upon is hashing. Non-cryptographic hash functions like perceptual hashing involve perceptually salient features of content, such as corners in images. They are more robust to changes that are irrelevant to how humans

perceive the content and aim to capture patterns that are relevant to semantic categories in a way that content remains identifiable even after perturbation.[13]

As an alternate mechanism, at the first stage, software such as Perspective API can be used to find out the probability of a piece of content being toxic. Then, at the second stage the human moderators can decide if the content is in reality offensive. This will mitigate the scope of human bias in the very first level of scrutiny. That is not to say that AI models are perfect. However, as is known, softwareare more likely to unlearn their biases as long as they have enough data.

To enhance the grievance redressal mechanism, content is censored or removed by the AI software as a first step. Thereafter, a redressal mechanism is provided to individuals to apply for review by a human moderator. Once the human moderator gives their decision regarding validity or non-validity of the content removed by the AI, the AI learns from the verdict and stores data regarding the verdict on the particular user's review request. From there on, if the same person applies for review again and their request is again rejected, the AI software can place the person's request in a lesser priority level with regards to dealing with complaints as the software will have learned that the particular person's content is more likely to be harmful. This will increase the efficiency of the redressal process by ensuring that users whose content has been repeatedly mistakenly taken down are provided priority during the appeal process.

The measures in place fail to effectively address the core issues at hand. In some countries this might be down to a lack of legislation, in others it may be down to a lack of public awareness or interest. An important tool to ensure that tech platforms dedicate resources to developing robust AI to tackle TFGBV is the promulgation of relevant laws. The government plays a critical role in addressing TFGBV. They can enact laws and sanctions to curb online violence. These laws should be comprehensive and cover a range of issues, including cyberstalking, revenge porn, and hate speech. Further, on a more practical level, individual liability should be assigned to users that engage in more serious forms of TFGBV such as cyberstalking, sextortion, revenge porn, deepfake porn, among others. However, it is imperative to understand here that the users alleged to have committed these crimes would be reported by the user(s) who were targeted online and not the tech platforms.

[13] *Ibid.*

As far as tech platforms are concerned, laws should require mandatory publishing of reports containing statistics and data regarding every step of the process, such as the content that was identified by the AI to be falling within TFGBV, the content that was removed, the content that went to the stage of human moderation, etc. Further, reports should also contain the statistics of users whose content has been repeatedly flagged and either removed or not removed.

Additionally, for users whose content has been repeatedly removed - their accounts should be reviewed by a human moderator after their content has been removed, let's assume 20 times. The human moderator is required to then analyse their content and decide if their account should continue to stay on the platform or blocked. Further, the same user should be blocked from creating additional accounts for a specific time period of like six months. This would ensure that users are held accountable for their content, so as to be cautious about the content they are posting.

The shortcomings of the introduced measures clearly show that there is not only a need to pass legislation but to also train law enforcement agencies to take digital violence more seriously and how to investigate/prosecute digital violence more effectively. To establish accountability for tech platforms, a centralised regulatory body should be established which analyses the reports submitted by tech platforms.

Further, tech companies should be mandated to report the content that was not removed from their platform. However, the most important aspect is the reporting on the users whose content was removed numerous times but the human moderator decided not to remove their account. The body is required to independently analyse their content, and report its findings to the tech platforms. It is understandable that this will increase workload for the tech platforms but in the overarching goal of societal interests, transparency is the need of the hour. Further, the body would merely be in advisory capacity. However, it would be responsible for reporting its recommendations to the government regarding users who should have been blocked but weren't. The government can then ask the platforms for reasons and explanations to ensure that platforms are following the regulations.

Summing up, it is relevant to clarify the role of social media companies, assign responsibility to the social media intermediaries, address structural inequalities, inform users/citizens about their rights to report content, formulate laws and establish enforcement mechanisms, support

independent research and evaluation of reporting that creates a deeper understanding of how social norms and sanctions are distributed in online communities.

5. FINAL CONSIDERATIONS

In conclusion, technology has added a new and concerning dimension to the issue of gender-based violence. While technology has the potential to empower women and promote gender equality, it also provides new avenues for violence and abuse. Therefore, it is crucial to address and prevent online gender-based violence and the harmful use of technology to protect individuals and promote gender equality. To this effect, the existing AI models need to be trained to overcome the existing gaps, laws are required to be formulated to clarify the rights and liabilities of users and platforms, and at last, enforcement bodies need to be established to ensure accountability of platforms.

THE CHALLENGES OF TACKLING EXTREMIST AND VIOLENT CONTENT IN THE PLATFORM GOVERNANCE FRAMEWORK

by

Andressa de Bittencourt Siqueira, Daniel Gadhof and Sophie Christiansen

1. INTRODUCTION

Content moderation measures are one of the biggest challenges in the platform governance realm, particularly when it comes to terrorist content given the extreme harms it poses, both online and offline. The dangers of extremism and radicalisation in online environments are widely acknowledged, especially with respect to the presence of terrorist organisations online.[1] The Internet can be misused by these groups or by individuals affiliated with them to intimidate, radicalise, recruit and facilitate the carrying out of terrorist attacks. The dangers of weaponising digital platforms for propagating extremism and violence reached the global centre stage with the Christchurch shootings being broadcasted live on Facebook. After the incident, the platform provider removed 1.5 million videos of the shooting globally in the 24 hours that followed the terrorist attack.[2] However, before its removal, the video was viewed 4000 times and even after the take-down of the video, it kept spreading across other social media platforms raising the classic whack-a-mole problem associated with content takedowns.[3]

It is clear that extremist and violent content online poses severe risks[4] to society that needs to be addressed by lawmakers and platform providers as key players in regulating it. However, attention is drawn to cases where there are state requests, ordering the removal of alleged terrorist content by platforms, opening doors for the pretext of illegitimate restrictions on the

[1]Douek E, 'Australia's" Abhorrent Violent Material" Law: Shouting" Nerd Harder" and Drowning Out Speech'(2020)' 94 ALJ 41.

[2] Chris Sonderby, 'Update on New Zealand' (*Meta*, 19 March 2019) <https://about.fb.com/news/2019/03/update-on-new-zealand/>

[3] Olivia Solon, 'Six Months after Christchurch Shootings, Videos of Attack Are Still on Facebook' *NBC News* (20 September 2019) <https://www.nbcnews.com/tech/tech-news/six-months-after-christchurch-shootings-videos-attack-are-still-facebook-n1056691>

[4] European Commission, 'Terrorist Content Online' <https://home-affairs.ec.europa.eu/policies/internal-security/counter-terrorism-and-radicalisation/prevention-radicalisation/terrorist-content-online_en>

right to freedom of expression. The question that remains is *to what extent* do the "extreme" harms of extremist and violent content justify imposing stricter content removal requirements on online platforms and *whether* there are differences in blocking uploads or removing it afterwards. In order to answer this question, first we analyse what exactly is meant by "extremist content"? Then we assess existing measures to tackle online extremist content and evaluate whether they are proportionate to the risks they pose. This is followed by a brief overview of the differences between the removal and the prevention of upload of extremist content.

2. CHALLENGES IN DEFINING EXTREMIST CONTENT

The concept of terrorism is highly debated due to its controversial reach and far-reaching implications. Definitions of terrorism are often problematic as the ambiguity of an overbroad definition gives opportunities for state censorship. On the other hand, by delimiting the concept too narrowly, the possibility of discussing the phenomenon and regulating it effectively may be impacted. In this section, given that tackling online terrorism is a common challenge among multiple actors involved in the Internet multistakeholder governance – state, supranational bodies, international bodies, companies, NGOs, users, technical community, and the academy –, we focus on the more recent attempts to define extremist and violent content.

Although the European Union Directive 2017/541, of 15 March 2017,[5] 'On Combating Terrorism" establishes that terrorist offences shall be defined in the national law, it provides some core criteria to better understand the phenomenon. Article 3 of the mentioned Directive states, as common ground, terrorist offences comprise intentional acts that, due to their nature or context, may seriously cause damage to a country or an international organisation (article 3.1).

It also sets a series of aims that lay down the characteristics of these offences, which are "(a) seriously intimidating a population; (b) unduly compelling a government or an international organisation to perform or abstain from performing any act; (c) seriously destabilising or destroying the fundamental political, constitutional, economic or social structures of a country or an international organisation" (article 3.2), as well as a list of nine specific terrorist

[5] Council Directive(EU) 2017/541 of 15 March 2017 on combating terrorism and replacing Council Framework Decision 2002/475/JHA and amending Council Decision 2005/671/JHA.

actions and the threat to commit any of those actions (article 3.1, points "a" to "j"). Journalistic, artistic and satirical content is expressly excluded from the scope of the application.

Regarding the concept of "extremist and violent content", the European Union Directive 2021/784 on Regulation Addressing Dissemination of Terrorist Content Online[6] (TCO Directive) states that it consists of material that incites the action, or the threat (e), of the actions listed on the points "a" to "i" of article 3.1 of the Directive on combating terrorism, in which there are, (a) directly or indirectly, the glorification or supporting of terrorist acts, (b and c) solicitation to commit, contribute or participate of these offences, (d) instructions regarding methods or techniques to commit or contribute to terrorist acts (article 2.7, points "a" to "e").

Another concept is brought by the Australian AVM Act (Abhorrent Violent Material Act), which was passed only a few days after the terrorist attack in Christchurch. The law creates a category of "Abhorrent Violent Material." The term, narrower than the one provided by the TCO Directive, relates to material that constitutes an act of terrorism, murder, attempted murder, torture, rape or kidnapping.[7] Crucially, however, content only falls under the definition if the material is recorded or streamed by the perpetrator or his accomplice.[8]

It is also important to note that platform providers, notably social media providers, also face some challenges in establishing the concept of terrorist content. For instance, on Facebook's Community Standards, one shall infer the scope of terrorist content from the sections related to "Violent and Graphic Content"[9] and "Hate Speech"[10]. In the mentioned platform, violent content (written or visual) targeting a person or a group is not allowed, while graphic (or explicit) content that involves violence when in the context of raising awareness about a sensitive situation is potentially permitted within Facebook, adopting in those cases a

[6] Council Regulation (EU) 2021/784 of 29 Aprile 2021 on addressing the dissemination of terrorist content online

[7] *Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019 (Cth) s 474.32(1)*

[8] Douek E, 'Australia's" Abhorrent Violent Material" Law: Shouting" Nerd Harder" and Drowning Out Speech'(2020)' 94 ALJ 41

[9] Meta,'Facebook Community Standards'<https://transparency.fb.com/policies/community-standards/violent-graphic-content/>

[10] Meta,'Facebook Community Standards'<https://transparency.fb.com/policies/community-standards/hate-speech/>

warning screen and not showing this kind of content to users under the age of 18. On Instagram Policies,[11] it is informed in its Help Center that the platform rules follow the TCO Directive and expressly mentions article 2.7.

The difficulties in establishing the concept of terrorism can also be found in the Oversight Board's performance. In the case "Video after Nigeria church attack"[12], analysed by Meta's Oversight Board, a graphic content showing the aftermath of a terrorist attack in June 2022 in Nigeria was posted with hashtags, whose aim was to raise awareness and document human rights abuse. The post was identified as violating Meta's rules for Instagram and removed by a so-called Media Matching Service bank ("escalations bank"). After the user appealed, a human reviewer still upheld the removal and, after a second appeal, this time to the Board, it reversed Meta's decision.

3. STRICTER CONTENT REMOVAL REQUIREMENTS

a) Measures against extremist and violent content

One can differentiate between three different sorts of measures against extremist and violent content: (i) first the platforms taking down user content on their own on the basis of violations of their community guidelines or local laws, (ii) then there are state takedown orders that oblige platforms to take down content and (iii) finally commitments under international collaborations between platforms, civil society and governments like Christchurch Call to eliminate terrorist and violent content online.

An example of the second group is the TCO Directive[13]. In order to face the dangers of extremist and violent content there are often stricter content removal requirements in legislative regulations. It states that authorities within the European Union can order service providers to remove extremist and violent content within one hour. There are exemptions for smaller platforms. If there are no current investigations, the platforms have to notify the person who has uploaded the content. The responsible person then has the opportunity to object against the removal. But the platforms are not obliged to scan content by themselves or use upload filters. They only have to remove content when there is a direct order from the

[11] Instagram, 'Dissemination of Terrorist Content Online' <https://help.instagram.com/548994106880972/?helpref=uf_share.>

[12] Oversight Board. Case 2022-011-IG-UA, 2022.

[13] Council Regulation (EU) 2021/784 of 29 Aprile 2021 on addressing the dissemination of terrorist content online

authorities. This is regardless of the place of the headquarters of the concerned company (deletion orders are possible across borders). Systematic violations will lead to fines of up to four per cent of annual revenue. The regulation will remain valid under the DSA, which provides a general approach for content moderation, even though not specifically for terrorism.[14]

In a very similar vein, the AVM also seeks to ban abhorrent material from platforms. According to the AVM, a platform provider is liable to prosecution if it learns of such content on its platform and fails to inform the police within a reasonable period of time. In addition, the platform provider is also liable to prosecution if it fails to take down violent content within a reasonable period of time. In both cases, the authorities can otherwise impose heavy penalties.[15] Regarding international agreements, Christchurch Call shall be highlighted. Even though it is a non-binding document, it establishes collective commitments by online platforms to tackle online terrorism.

b) Proportionality

At first it needs to be said that extremist and violent content is not protected by freedom of expression and the Internet is not a lawless zone.[16] Nevertheless, measures taken to prevent the extreme harms of extremist and violent content need to be proportional. There is the risk of overblocking or overremoval of legitimate content as removing objected content is the cheapest and safest way to avoid liability.[17] There is the danger that political speech, religious speech or news reporting may be silenced. In that way, it could also lead to a distortion of important political conversations.[18] This could also lead to echo chambers and chilling effects on public conversations. In cases and systems where claims about extremist content come

[14] European Commission, 'Fragen und Antworten: GesetzüberdigitaleDienste' (*European Commission - European Commission*, 14 November 2022) <https://ec.europa.eu/commission/presscorner/detail/de/qanda_20_2348>

[15] Douek E, 'Australia's" Abhorrent Violent Material" Law: Shouting" Nerd Harder" and Drowning Out Speech'(2020)' 94 ALJ 41.

[16] EU2020,'Fight against terrorism on the internet: Political agreement reached on EU Regulation' (11 December 2020)<https://www.eu2020.de/eu2020-en/news/pressemitteilungen/fight-terrorism-internet-eu2020/2426232.>.

[17] Keller D, 'Internet Platforms: Observations on Speech, Danger, and Money' [2018] Hoover Institution's Aegis Paper Series.

[18] Douek E, 'Australia's" Abhorrent Violent Material" Law: Shouting" Nerd Harder" and Drowning Out Speech'(2020)' 94 ALJ 41.

from other users or algorithmic filtering techniques of platforms themselves, it must be taken into account that the platforms have little expertise in the legal assessment of content and balancing competing rights, much less in several local languages and regional policies. This challenge becomes more exacerbated in shorter takedown timeframes.

Thus, while shorter timeframes for content takedowns are helpful to quickly and urgently reduce the harm that can be caused by extremist content online they also pose the risk of chilling free speech[19]. There is a tradeoff between the interest of taking down extremist content as quickly as possible and effective protection of freedom of speech of the users from their content being wrongfully removed because of the platforms aim to avoid liability.

Another grave danger that needs consideration is the risk of pretext: nation-states are able to order platforms to remove specific content. When there are state requests, ordering the removal of alleged extremist and violent content by platforms, this could open doors for the pretext of illegitimate restrictions on the right to freedom of expression.

For example, Russia removed satirical content as "terroristic".[20] This is especially problematic as every country has this opportunity, including those who do not espouse democratic values. This could be used in order to remove content they do not agree with under the excuse of "terrorism"[21], especially when a country is facing internal turmoil or warfare. Thus, in order to protect free expression rights and also the right to a fair trial, due process and the possibility for remedy for the user need to be provided. Platforms must provide users with reasons for content removal and inform them of their right to legal recourse.

The singular focus on the internet and overreliance on content purges as tools against real-world violence could miss out on or even undermine other interventions and policing efforts.[22] Moderate voices, who share experiences and grievances with potential extremists

[19] ibid

[20] Keller D, 'Internet Platforms: Observations on Speech, Danger, and Money' [2018] Hoover Institution's Aegis Paper Series

[21] Tomas Rudl, 'KeineUploadfilter-Pflicht: EU einigtsich auf GesetzgegenterroristischeInhalteimNetz' (*netzpolitik.org,* 10 December 2020) <https://netzpolitik.org/2020/keine-uploadfilter-pflicht-eu-einigt-sich-auf-gesetz-gegen-terroristische-inhalte-im-netz/>

[22] Keller D, 'Internet Platforms: Observations on Speech, Danger, and Money' [2018] Hoover Institution's Aegis Paper Series

but who oppose violence are an important tool in the fight against terrorism. When these people are silenced their beneficial effect is barred and mistrust, anger and frustration with the government could arise. In this regard, content removals could lead to the opposite of the pursued goal as the efforts may cultivate precisely the attitudes and animosities that counter-radicalization efforts are supposed to prevent by creating feelings of alienation and social exclusion.[23] Furthermore, by removing extremist and violent content there is a restriction of valuable sources of intelligence in the fight against terrorism and of evidence of war crimes or other atrocities.[24]

4. DIFFERENCES BETWEEN REMOVAL AND PREVENTION OF UPLOAD

There are different outcomes in removing content after its posting and preventing it from being uploaded at all. Removing content that is subsequently found to be violent or terrorist is a much less far-reaching interference with the user's freedom of expression. So when it comes to blocking content identified as terrorist from being uploaded, there are some aspects that are worth analysing with caution.

The database of hashes managed by the Global Internet Forum to Counter Terrorism shall be highlighted. It is shared among many platforms to avoid the duplication of terrorist images and videos in online environments. In this mechanism, a unique signature is given to a certain video or image that allows its identification, and consequent removal or prevention of publication, if the same content is uploaded again on any of the participant platforms. Most platforms perform an *ex-post* removal, in which the duplicated content is only checked after its posting. Some platforms, such as YouTube, carry out the filtering before making the content public.

However, the automatic identification of terrorist content is a problem in both *ex ante* and *ex post* removal. Although, it looks flawless in theory, the application of the hash filters is not immune to controversy. Most platform providers modify all uploaded images, by e.g. resizing or compressing. For this reason, even though the images are similar, the hash is different from one another, potentially jeopardising the automatic identification[25]. Another issue

[23] Douek E, 'Australia's" Abhorrent Violent Material" Law: Shouting" Nerd Harder" and Drowning Out Speech'(2020)' 94 ALJ 41.

[24] ibid.

[25]Farid H, 'Reining in Online Abuses' (2018) 19 Technology & Innovation 593.

regarding the use of hashes – but not limited to it – is the lack of context, since the technology alone cannot correctly identify e.g. if the user's post or comment is raising awareness, whistleblowing or even asking for help. Due to the large amount of content that is posted on a daily basis, it would not be possible for human moderators to check the content for legitimacy. Therefore, content would have to be checked by machine, for example by Artificial Intelligence (hereinafter AI). The risk that the AI will incorrectly classify the content as terrorist is high, especially for content that deals with this topic in an educational way (e.g. news), since the AI can take into account what is actually written and being displayed and does not "read between the lines". The main problem regarding extremist and violent content is thus the removal without human review.

In addition, with an upload block, there is always the risk of pre-censorship, since third parties already from the outset do not have the opportunity to see the content. Upload blocking thus represents a significantly stronger intervention. This is not ruled out in principle, but the requirements, for example with regard to transparency, would have to be significantly higher in order to protect users in their right to freedom of expression.

5. CONCLUSION

All in all, moderating content – especially extremist and violent content – is no easy task in the realm of platform governance given the scale and speed at which it must operate. Each of the players shall take actions to contribute to the prevention of the spread of online terrorism[26]. Recognized as gatekeepers[27], custodians[28] and new governors[29], platforms have increasingly started to accept responsibility to prevent the problems that arise out of extremist content.

However, there is a greater need for ensuring transparency to the user who has his or her content moderated, notably through due process, as well as accountability in the case of abuse

[26] Based on the concept of platform governance provided in the World Summit on the Information Society (WSIS) and debated in the book Kurbalija, Jovan. An Introduction to Internet Governance. 7th ed. DiploFoundation, 2016. p. 5-6.

[27] Celeste E, 'Digital Constitutionalism: A New Systematic Theorisation' (2019) 33 International Review of Law, Computers & Technology 76

[28] Gillespie T, 'Custodians of the internet – Platforms, content moderation, and the hidden decisions that shape social media'(Yale University Press, 2018) p. 209.

[29] Klonick K, 'The New Governors: The People, Rules, and Processes Governing Online Speech' (2017) 131 Harv. L. Rev. 1598

of power by platforms and supervision of their actions. The user's right to free expression needs to find sufficient consideration and the measures taken have to be proportional in regard to their actual efficiency.

ABOUT THE NATIONAL LAW UNIVERSITY DELHI (NLUD)

The National Law University Delhi is one of the leading law universities in the capital city of India. Established in 2008 (by Act. No. 1 of 2009), the University is ranked second in the National Institutional Ranking Framework for the last five years. Dynamic in vision and robust in commitment, the University has shown terrific promise to become a world class-institution in a very short span of time. It follows a mandate to transform and redefine the process of legal education. The primary mission of the University is to create lawyers who will be professionally competent, technically sound, and socially relevant, and will not only enter the Bar and the Bench but also be equipped to address the imperatives of the new millennium and uphold constitutional values. The University aims to evolve and impart comprehensive inter-disciplinary legal education which will promote legal and ethical values, while fostering the rule of law.

The University offers a five year integrated B.A., LL.B (Hons.) and one-year postgraduate masters in law (LL.M.) along with professional programs, diplomas and certificate courses for both lawyers and non-lawyers. The University has made tremendous contributions to public discourse on law through pedagogy and research. Over the last decade, the University has established many specialised research centres including the Centre for Communication Governance (CCG), Centre for Innovation, Intellectual Property and Competition, Centre for Corporate Law and Governance, Centre for Criminology and Victimology, and Project 39A. The University has made submissions, recommendations, and worked in advisory/consultant capacities with government entities, universities in India and abroad, think tanks, private sector organisations, and international organisations. The University works in collaboration with other international universities on various projects and has established MoU's with several other academic institutions.

ABOUT THE CENTRE FOR COMMUNICATION GOVERNANCE

The Centre for Communication Governance at the National Law University Delhi (CCG) was established in 2013 to ensure that Indian legal education establishments engage more meaningfully with information technology law and policy and contribute to improved governance and policy making. CCG is the only academic research centre dedicated to undertaking rigorous academic research in India on information technology law and policy in India and in a short span of time has become a leading institution in Asia. Through its academic and policy research, CCG engages meaningfully with policy making in India by participating in public consultations, contributing to parliamentary committees and other consultation groups, and holding seminars, courses and workshops for capacity building of different stakeholders in the technology law and policy domain. CCG has built an extensive network and works with a range of international academic institutions and policy organisations. These include the United Nations Development Programme, Law Commission of India, NITI Aayog, various Indian government ministries and regulators, International Telecommunications Union, UNGA WSIS, Paris Call, Berkman Klein Center for Internet and Society at Harvard University, the Center for Internet and Society at Stanford University, Columbia University's Global Freedom of Expression and Information Jurisprudence Project, the Hans Bredow Institute at the University of Hamburg, the Programme in Comparative Media Law and Policy at the University of Oxford, the Annenberg School for Communication at the University of Pennsylvania, the Singapore Management University's Centre for AI and Data Governance, and the Tech Policy Design Centre at the Australian National University.

The Centre has had multiple publications over the years including reports on Intermediary Liability in India, a report Mapping the Blockchain Ecosystem in India and Australia, an authored UNDP Guide on Drafting Data Protection Legislation, a book on Privacy and the Indian Supreme Court, Hate Speech Report, and most recently two essay series, one on Democracy in the Shadow of Big and Emerging Tech, and a second on Emerging Trends in Data Governance. The Centre has launched freely accessible online databases - Privacy Law Library (PLL) and High Court Tracker (HCT) to track privacy jurisprudence across the country and more than sixteen jurisdictions across the globe in order to help researchers and other interested stakeholders learn more about privacy regulation and case law. CCG also has an online 'Teaching and Learning Resource' database for sharing research oriented reading

references on information technology law and policy. In recent times, the Centre has also offered courses on AI Law and Policy, Technology and Policy, and First Principles of Cybersecurity. These databases and courses are designed to help students, professionals, and academicians build capacity and ensure their nuanced engagement with the dynamic space of existing and emerging technology and cyberspace, their implications for society, and their regulation. Additionally, CCG organises an annual International Summer School in collaboration with the Hans Bredow Institute and the Faculty of Law at the University of Hamburg in collaboration with the UNESCO Chair on Freedom of Communication at the University of Hamburg, Institute for Technology and Society of Rio de Janeiro (ITS Rio) and the Global Network of Internet and Society Research on contemporary issues of information law and policy.

Centre for Communication Governance at National Law University Delhi, Sector 14, Dwarka, New Delhi, 110078, India

ccgdelhi.org | **@CCGNLUD**
Email: ccg@nludelhi.ac